UNIVERSIDADE D
COIMBRA

Alban Emmanuel MAZARS-SIMON

THE WILD IN LIVE PROJECT:

A HUMAN/ALGORITHM LEARNING NETWORK TO HELP CITIZEN SCIENCE IN WILDLIFE CONSERVATION

Dissertação no âmbito do Mestrado em Ecologia orientada o Professor Doutor José Paulo Filipe Afonso de Sousa e apresentada ao Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Junho de 2019

Alban Emmanuel MAZARS-SIMON

The Wild in Live Project: A Human/Algorithm learning network to help citizen science in wildlife conservation

Dissertação no âmbito do Mestrado em Ecologia orientada o Professor Doutor José Paulo Filipe Afonso de Sousa e apresentada ao Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Junho de 2019

Abstract

Despite the recent advances, citizen science is limited by multiple biases specific to human monitoring that can hamper the quality of the dataset. However, the advent of new technologies is revolutionising the generation and access to conservation data. The Wild in Live project aimed to prove if such technology could be used in citizen science as a tool to help the data collection. We, therefore, created an algorithm capable of scouting social media platforms to categorise and detect sea turtles. The prototype has a classification accuracy of 95.53% and can recognise individual sea turtles within a considerable database of 22 500 pictures. The preliminary result offered a good insight into the untapped potential and limitation of computer-vision to monitor ecosystems. The passive collection of data will increase the scope and scale of sea turtle monitoring, and Wild in Live project could become a tool to help researchers gather more information while citizen science projects could use it to raise awareness.

Keywords

Machine Learning, Social Media, Citizen Science, Image Classification, Conservation Biology.

Resumo

Apesar dos recentes avanços, algumas das iniciativas de ciência cidadã estão ainda limitadas por múltiplos constrangimentos específicos à monitorização efectuada pelas pessoas que participam deste tipo de projectos e que podem impedir a qualidade dos dados recolhidos. No entanto, o advento de novas tecnologias está a revolucionar a recolha e o acesso a dados de conservação. O projeto Wild in Live teve como objetivo provar se tal tecnologia poderia ser usada em ciência cidadã como uma ferramenta para auxiliar na recolha de dados. Foi desnevolvido um algoritmo capaz de rastrear plataformas de redes sociais para categorizar e detectar tartarugas marinhas. O protótipo tem uma classificação com uma precisão de 95,53% e pode reconhecer tartarugas marinhas individuais dentro de um banco de dados considerável de 22 500 imagens. O resultado preliminar ofereceu uma boa visão do potencial inexplorado e da limitação da visão computacional para monitorar os ecossistemas. A recolha passiva de dados aumentará o escopo e a escala do monitorização de tartarugas marinhas, e o Wild in Live poderá tornar-se uma ferramenta para ajudar os pesquisadores a recolher mais informações, enquanto projetos de ciência cidadã poderiam usá-los para aumentar a conscientização para a conservação das espécies.

Palavras-chaves

Machine Learning, Redes Sociais, Citizen Science, Classificação de Imagens, Biologia da Conservação.

Acknowledgements

During the development of this project, I have received a great deal of support and assistance. I would first like to thank my colleagues, who helped to develop the project, Antoñio Alves, believed in our ideas and allowed us to create something amazing. I want to acknowledge my equal, Joao Amorim at Critical for his excellent collaboration. I would particularly like to thanks the ISTS researchers for their much-appreciated feedback on the prototype and for all of the opportunities I was given to further my project. I would also like to thank the head of the AI department at critical, Paulo Gomes, for his guidance. Who provided us with the tools that we needed to choose the right direction and complete the prototype of Wild in Live. Also, I would like to thank my brother for his wise counsel and a sympathetic ear to start this project. Moreover, my supervisor Jose Paulo Sousa, for his rapid but fantastic feedbacks.

Finally, My grandfather, whose support exceeded was beyond anything I could ever ask for. You were the reason why I made this far. Thank you.

Table of Contents

List of abbreviations

AI: Artificial Intelligence

API: Application Programming Interface

CNN: Convolutional Neural Network

DBI: Database and Interaction

I³S: Interactive Individual Identification System

IMA: Image Processing Algorithm

ML: Machine Learning

SIFT: Scale Invariant Feature Transform

List of Figures

List of Tables

INTRODUCTORY SECTION – putting this thesis into context

The need for citizen science

Living ecosystems are constantly changing, always subject to an endless possibility of disruptions, allowing new species to take place in the communities. Hurricanes, volcanoes, shifting continents are forces that are famous for redistributing earth fauna and flora (Fattorini *et al.,* 2018). Humans are as well, considered as a force for ecological change and one to be reckoned with. Alteration of natural habitats by humanmade perturbations are known to cause broad changes in the ecosystem wellbeing (Madin *et al.,* 2013), and its influence is felt through the different ecological scales (Atkinson & Cameron, 1993, Hendry *et al.,* 2017, Madin *et al.,* 2013).

This influence was in part, translated into an increase background extinction rate, which the magnitude makes it both exciting and depressing to witness. If the recent loss of species is significant, it does not yet qualify as a mass extinction event in the paleontological sense; however, there is a concern that with the current extinction trajectory we could be within a few human generations of a mass event (Matzke *et al.,* 2011). So here lies one of the main difficulties that encounter species today, the change in the environment is too fast to leave any chances to adapt appropriately. Under the current scenario about 20 % of all species are projected to become extinct within 30 years(Singh, 2002), with 25 % of mammal species declining (Schipper *et al.,* 2008), 12% of bird species, and with the amphibians that sustain most of the loss of the vertebrates with 43.2% of their species currently declining rapidly (Young *et al.,* 2004).

Given that most of the essential functional aspects of the ecosystem are depending on biotic interactions, the loss of biodiversity has an unwelcome negative influence on ecosystems functions and provoke the acceleration of species local extinction (Cocucci *et al.,* 2014). The adverse effects also appear to have an increasing impact on species which have an essential role in the ecosystem and coupled with a diminished diversity interaction (Hendry *et al.,* 2017), it has caused an impoverishment of ecosystem services (Bullock *et al.,* 2011).

Determining the current state of the ecosystems function and understanding which ones are at risk of not being able to function correctly is crucial for the prioritisation of conservation and restoration efforts. However, long term trends in species occurrences have faced difficulties as the lack of data and robust methodology hamper the determination of the differences between the occurrence of opportunist or functional species and what species individual contribution to the ecosystem functions (Oliver *et al.,* 2015). However, despite the efforts of ecologists and conservationists to protect threatened communities, their ecological extinction may have already had happened (Cocucci *et al.,* 2014).

Changes in the environment do not often affect biodiversity immediately. Delayed responses are still poorly understood, and consequently, the full scope of consequences due to rapid environmental changes are often underestimated as it is difficult to contextualise the implications (Essl *et al.,* 2015).

The pervasive effect is accentuated with perception issues called a “shifting baseline”. As Humans, we always update our perception unconsciously to our surroundings while giving more attention to the latest changes and paying less attention to the previous state, and therefore, for us, the abnormal becomes the new normal (Papworth *et al.,* 2009). Generational amnesia can be a significant problem while setting conservation goals as the target set may not be representative of the original environmental state if no data are available to prove the contrary (Danielsen *et al.,* 2000). Understanding the scale of time must be improved to consider the mechanisms causing time lags and shifting human perception, and notably by improving and expanding existing long-term ecological monitoring (Essl *et al.,* 2015).

Long term ecological monitoring is valuable for many aspects, as it can help to evaluate responses to changes to a complex system such as climate change and ecosystems disturbance and offer insights by providing the baseline to appreciate changes. Detecting and evaluating the responses of ecological function as a response to conservation management intervention (Lindenmayer & Likens, 2009).

Monitoring can differ in techniques, approaches and focus. Commodity-based monitoring will focus on economic matters such as monitoring exploitable resources like fisheries (Dearden & Lunn, 2006) and forestry (Tang & Shao, 2015) but more recently, the focus has also shifted to include ecological and social responses as well.

Non-commodity monitoring will, however, look at issues that do not seem to be economically significant such as monitoring an indicator species (Hendry *et al.,* 2017), songbirds call (Buxton *et al.,* 2016) or water quality (Behmel *et al.,* 2016).

Monitoring activities will look at different aspects of the ecosystems such as composition, structure or processes and also include different types of assessment of ecosystems;

- Status (i.e. monitoring of population),
- Impact (i.e. pollution assessment),
- Adaptive management (Conrad & Hilchey, 2011).

In 1998, the United Nation Economic Commission for Europe (UNECE) adopted the Aarhus Convention, which established several rights for the public in regard for the environment. Such as the right to access environmental information, the right for public participation in environmental decision making and the right to review procedures to challenge decisions that have been made with respecting the two previous rights (Aarhus Convention, 1998).

The Aarhus Convention helped to establish the framework of different type of monitoring governance and engage the public into such activities (Mason, 2010).

a) ***Consultative governance*** suggests that a government or a central agency is asking for information directly from the public, a form of participation qualified as “top-down” with the purpose to provide early detection of issues or environmental concerns by citizen which can then be investigated by the scientific community (Conrad & Hilchey, 2011). This form of monitoring has advantages to be used in an area where there is illegal poaching of protected species (especially in developing countries), the citizens could serve as a “watchdog” assistance for government or conservationists as the data collected could be used to create long term dataset to be used by scientists (Whitelaw *et al.,* 2003).

b) ***Collaborative governance*** often represents multiple facets of a community with business, government and stakeholders. With its collaborative nature, it often yields more decision power than other types of monitoring, as every stakeholder is represented equally within the project (Whitelaw *et al.,* 2003).

c) ***Transformative governance*** describes groups being managed from the "bottom up" are often born out of a crisis and bear the goal to initiate government action on specific and local issues, with initiatives, funding and leadership provided by the local community (Conrad & Hilchey, 2011). In opposition to consultative governance, this model has the advantage to involve participants in every stage of the monitoring project, from defining the issues through the communication of the results and taking actions. Scientist and researcher role in transformative governance is to advise and guide the community.

Monitoring allows for citizens to be involved in science as "researchers" (Linda E. Kruger, 2002). Citizen science includes community based monitoring which is a management where citizen and stakeholders are included in the management of a natural resource (Keough & Blahna, 2006) and/or "a process where concerned citizens, government agencies, industry, academia, community groups, and local institution collaborate to monitor, track and respond to issues of common community concern" (Whitelaw *et al.,* 2003)

The different type of governances will be more effective in specific citizen science project (and communities), with transformative and collaborative governance being more commonly associated with smaller scale participation while consultative governance being more adapted for larger geographic scales (Conrad & Hilchey, 2011).

I. The origins of Citizen Science

Before the 19th century, what was then defined as science was carried by amateurs. Michael Faraday (1791-1867) for example, who discovered the principles underlying electromagnetism induction, electrolysis and diamagnetism, never had any formal scientific education, referring as himself as an "experimental philosopher" to the end of his carried as director of laboratories of the Royal Institution in London (Secord, 2005). When considering the modern work of citizen science, one should remember that the word "scientist" was only coined in 1833 by William Whewell which has led to the professionalization of science and excluded citizens (Ross, 1962) The modern approach to science is challenging a century-old approach that science is only to be conducted by experts. In 1995, Alan Irwin coined the term "citizen science" to describe expertise that exists among those who are traditionally seen as ignorant "lay people", and should be used as complementary due to the uncertainty that they could represent (Irwin, 1995).

Another renown figure in citizen science, the American researcher of the Cornell Lab of Ornithology, Rick Bonney, has a different interpretation on the public participation in science.

He compares citizen science as a crowd founding practice as it defines to outsource a job traditionally held by an agent to an undefined, large group of people in the form of an open call (Science Communication Unit, 2013).

Humans have always recorded and monitored our environment to better understand our surrounding, like some of the oldest record of a phenological event such as the timing of cherry blossom in Kyoto for 1200 years (**Figure 1**) which have been used for climate analysis (Aono & Kazui, 2008). However, one of the longest-running crowd-sourced science projects is the Christmas Bird Counts by the National Audubon Society which began surveying wintering bird population in 1900 with no interruption to this day (Kobori *et al.,* 2016). This survey is open to everyone, even with no experience nor knowledge in birdwatching as long that they are situated in the study areas. In 2019, just under 77,000 field observers participated in identifying almost 60 million individuals in 2673 species (Chandler Lennon, 2019). The results are then distributed online, and they help conservationist to study the health and status of bird from the North American population.

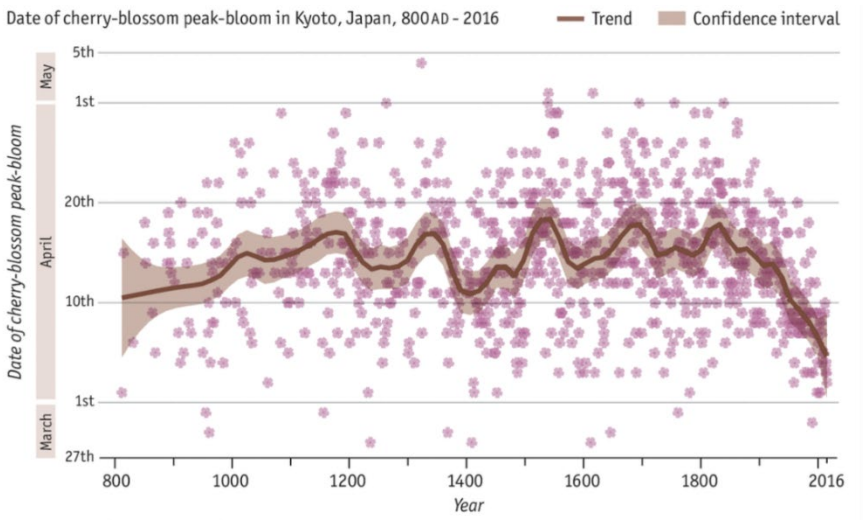

Figure 1: 1200 years of monitoring of the blossom of the cherry tree in Kyoto. In recent years, the blossom starts in late March due to climate change. Credit Yasuyuki Aono, Osaka Prefecture

During the last decades, the advent of new communication technologies has made it possible for amateurs to enlist in extensive participatory studies providing scientific data with an emphasis on scientifically sound and measurable goal for public education (Science Communication Unit, 2013). Different projects were created, each encompassing different methods or models of citizen science with a wide range of subjects (i.e. Zooniverse for astronomy; Old Weather for climatology; Seafloor Explorer for ocean exploration) (Citizen Science Alliance, 2011).

II. Advantages of Citizen Science

By enabling people to be invested in a scientific project, citizen science can contribute to firstly provide the opportunity to create new knowledge and insights relevant to science or management and government and secondly, it will allow individuals to acquire knowledge, skills and gain scientific literacy through their involvement in citizen science projects (Bela *et al.,* 2016).

Thirdly, it will empower the individual through its active civic participation in the project, allowing them to take part in policy debates and decision-making processes. Scientific participation is the opportunity for a citizen to be in the centre of knowledge creation, civic participation, and most importantly learning about the situation (Turrini *et al.,* 2018).

One of the critical aspects of citizen science is the generation of new knowledge while addressing scientific objectives which makes citizen science projects remarkably successful in generating an impressive quantity of data, especially spanning over a large spatial area and temporal extents (Devictor *et al.,* 2010).

This allows the data to be the basis for an analysis of trends and environmental changes while supporting local and international monitoring, planning and administration of the environment (Turrini *et al.,* 2018). The development of collaborative initiatives between scientists and members of the public offers the opportunities and means necessary for in-depth learning (Bela *et al.,* 2016), as citizen science is usually aimed to encourage such endeavour, it increases member's awareness of the issues while gaining environmental insight (Johnson *et al.,* 2014). The democratisation of science in conservation biology include more multidisciplinary topics to capture the attention of the policymakers. As citizen science is an effective way to gain the participation of stakeholders and the general public into the planning and management, the community-based monitoring tends to be more engaged with the issue and participate more intensely on the development and management of the project which increase the project' success rate (Jarvis *et al.,* 2015). The success can be attributed to the social capital as the public support for the project increase, and it helps to establish trust and cooperation with the community involved in the monitoring project, as it can lead to volunteer engagement, problem-solving, resource identification and agency networking (Conrad & Hilchey, 2011).

Citizen science offers a non-negligible alternative to government agencies as its implementation and management are significantly more cost-effective than direct government monitoring as the fieldwork can spawn over large areas and often during regular office hours (Whitelaw *et al.,* 2003).

The diversity offered by citizen science allows researchers to collect data on the environment or ecosystems that are not commonly investigated, with distant places or the inability to be physically present directly on the site can hamper with the scientist ability to have continuous monitoring (Evans *et al.,* 2005).

III. The challenges of Citizen Science

Scientific works have been carried for the last centuries by “experts”, people with a profound knowledge of their subject and as such, many experts have questioned the quality and reliability of data collection conducted by amateurs and often been regarded as substandard or doubtful (Engel & Voshell, 2002). So perhaps, here lies one of the substantial challenges facing crowdsourced data collection, getting the trust and recognition of the scientific community and any potential users of citizen science gather data. Which is an understandable stance as large datasets due to their sampling design, be flawed by bias or pitfalls (Lepczyk, 2005).

1. Random errors

In any ecological study, one of the main goals is to determine if the variations in response to an interference are due to predictors. Also, random errors represent the variation in a response that cannot be imputable to specifics predictors. In the context of crowdsourced monitoring, sampling error is more often due to the ability of each observer to detect, identify and estimate the subject monitored. Mistake and errors can easily be introduced through recording covariate data or due to the variability of the execution of the monitoring protocol. If those errors are not accounted for in a model, they are then included in the overall random error, which may complicate the detectability of the trends studied (Bird *et al.,* 2014).

2. Bias

Different biases are known to be explicitly present in citizen science monitoring, one of which is related to plant or animal surveys and is called “heterogeneity of species detection” (Devictor *et al.,* 2010). This is the systematic bias; it happens when repeated measure with the same method provides a constant under or overestimate of a fixed value.

This bias is inherent to the ability of the observer to actually detect the studied subject, as the counts are shaped by the appearance or absence of the subject. The result may display an unknown variation in the abundance of individual studied id the variation in detectability is not accounted for (Royle *et al.,* 2007).

In 2012, Farmer, Leonard and Horn's study highlighted the importance of testing observers' skills in data collection and modelling not to encourage the use of opinion-based indications of observer confidence (Farmer *et al.,* 2012).

This is the overconfidence bias, and it occurs when moderately competent observers were significantly less likely to report false positive records of rare species than common species, while experts were significantly more likely to report false positives of rare species than common species. Because false positives can also result from observers overestimating their own skills.

Another example of measurement bias happens when divers must estimate the size of a fish. Usually, the sizes of large individuals are overestimated whereas the smaller individuals are underestimated, typically due to magnification and other factor affecting the observer perception while underwater (Edgar *et al.,* 2004)

When some aspects of the process of interest seem to be more sampled than others, we might be in the presence of a sampling bias which will have a significant change in the statistical interpretations as the mean will be overly influenced by these samples (Bird *et al.,* 2014).

Another common bias for dataset when collected by multiples observers, is the variability that arises between them in their sampling efficacity. The mean value of the measurement described by the volunteers may be close to the real value, but some volunteers may contribute more sample than others and therefore skewing the values collected. If the observations made by a single volunteer are consistently underestimated or overestimated, then by considering each observation as an independent sample could have the potential to bias the entire estimate of a trend of a mean (Bird *et al.,* 2014).

The volunteer effort may also vary due to declining commitment and availability, the weekend bias describes an overabundance of data relative to the week and may underestimate the presences of species and making more difficult to distinguish between seasonal patterns over a more extended period (Courter *et al.,* 2013).

Citizen science will only be available to areas accessible to the crowdsourcing community. Therefore, explaining the overabundance of terrestrial monitoring projects as it usually faces fewer challenges than projects in marine settings.

Aside from all the logistical challenges and safety measures that a marine project will demand, it may be more challenging to find recruit and train volunteers willing to accompany scientist for those gruelling and lengthy monitoring sessions where it may be harder for them to “take ownership” of a site without visible demarcation nor boundaries (Cigliano *et al.,* 2015).

3. Difficulty to share the data

The data collected by crowdsourced monitoring needs to be openly shared with both the participants and thirds parties, as it improves the data relevance and quality for science and policy management.

The only way for the records to be exploitable as ecological data, they must be submitted in a defined standardised format that has been objectively selected to be included and stored in the corresponding database (Ganzevoort *et al.,* 2017). However, one must be wary of the tensions that may arise when the data collected by volunteers become too dissembled from contextual richness lost during the process; as it has direct effects on both the data collection the motivation of volunteers. If there is a lag time between the monitoring event and the publication of the study in academic review, it can be challenging for the volunteers of the project as the participants may have to wait years before seeing the results of the study (Tenopir *et al.,* 2011).

It requires scientists and managers of a citizen science project to listen and consider the views of the volunteers regarding ownership and the appropriate use of their data (Turnhout & Boonman-Berson, 2011).

IV. Mitigating the challenges in Citizen Science

1. Project design and volunteer’s motivation

With crowd-sourced monitoring, there is an underlying assumption that as a source of knowledge, the project requires the volunteers to conform to the scientific method. However, more flexible engagements can help curb the requirement by allowing the volunteers to have more influence on the project design and empowering them to exchange with the other actors among which scientist are only one of the kinds (Hecker *et al.,* 2018). Those are frequently addressed with face to face interactions to increase trust and inducing a real sense of transparency. By supporting those collaborations, citizen science project can benefit from creating and modifying techniques or processes, reducing information asymmetry and cognitive complexity while increasing transparency through the different views of the stakeholder (Novak, 2009).

Those approaches still come with substantial time investment for the managers of the project; however, it allows to get a better grasp of the full range of factors that motivate volunteers.

Such factors can emanate from a desire to discover and learn new things to further personal interest or a yearning to help the project as the volunteer share values and belief that come with such project (e.g. free and shared knowledge). Volunteering can also be motivated by the human desire to meet new people sharing the same interest, to gain recognition or solely for divertissement (Hecker *et al.,* 2018). It is also seen as a way to gain experience during the project for the future career of the participant (Geoghegan *et al.,* 2016). All those engagements except for career development are intrinsic motivations.

To ensure engagement from the volunteers, regular feedbacks and recognition are essential to maintain motivation (Rotman *et al.,* 2012). The interactions between scientists and volunteers are essential for creating social cohesion through regular group activities and regular face to face meetings but require effort from both parties. Moreover, even if spatial constraints can be mitigated through online interactions, it tends not to be the most representative as usually, a few volunteers provide most of the contribution, while others are considered as passive during the project with only a few punctually participating (Hecker *et al.,* 2018).

This is also better known as the 90-9-1 rule coined by in 2006 by Nielson while describing that “in most online communities, 90% of users never contribute, 9% of users contribute a little, and 1% of users account for almost all the action” (Nielson, 2006).

Unfortunately, participation inequality will be always present regardless of the actions taken in every online community and user services. However, it is possible to achieve more equitable distribution (e.g. 75-20-5), by taking steps to help increase participation engagement:

- Facilitating contributions: allowing the user to give input with minimal effort required such as star rating and minimalising natural language inputs.
- Participation side effect: collecting data from other activities realised by the user on the service that is exploitable as an input.
- Gentler learning curve: letting the user learn the service by allowing to modify but not create from scratch input. By building their contribution, the user will be more compelled to complete than instead, recreate entire entities.
- Fair reward system: reward is a useful enticing tool to motivate user from the punctual base to a more active role while cementing active users. Perks and recognition promoting quality contributors who have proven value (e.g. reputation ranking).

Therefore, a successful participation model must include simple tasks with little complexity or cognitive effort (e.g. monitoring) along with more complex activities which require more significant commitment (e.g. data analysis). Face to face meetings or workshops and design-creation can be combined with online participation with different options according to the different knowledge required with a flexible continuity in the engagement. Engagement can be more easily created with game design element incorporated to the project to reward achievement of volunteers and allowing for social recognition and reputation that are combined with regular feedback from the project (Bowser *et al.,* 2013). Sharing issues and solutions with the volunteers of the project allow to alleviate information asymmetry and facilitate the application of new tools.

2. Random error fix

If a dataset contains a large number of random errors hampering the ability to detect patterns and distinguish trends, it may not necessarily be a problem if an increased quantity of crowd-sourced data can be collected to offset the issues (Bird *et al.,* 2014). However, in cases when no additional data can be collected, accounting for random errors in both metadata and covariate is required. Identifying the overall effectiveness of a volunteer by attributing them an identifier to relate different metrics (e.g. engagement, training and experience) to determine better how sampling was performed and to help understand variations (Snäll *et al.,* 2011). To be able to measure the effort made by each volunteer is an essential factor to be integrated into standardised monitoring data.

However, covariates can include factors that can have an impact on the data collected (i.e. underwater visibility for visual survey affect the observer regardless of its skills) (Edgar *et al.,* 2004).

Modelling approaches can offer options that account for the different types of errors and biases encountered in the dataset.

3. Linear Models

One of the most widely used statistical tools to quantify random errors in ecological data are the linear models. The premise of the linear models is that the changes in the response data could be the results of studied predictors as a linear function, covariate or metadata which are described as fixed effects. A linear model can be extended by allowing non-linear relationships between the predictor and response data with the use of smoothing functions with multiple degrees of freedom as an additive model (Hastie & Tibshirani, 2007).

Usually, a significant amount of variations in the response data can be described simply with relationships. However, it is unlikely that the only available predictors and covariates are the only factors influencing the response variable. Therefore, any variations can be regarded as a stochastic process if it cannot be accounted for used parameters and can be described using a probability distribution, which allows the remaining or residual variations to be described based on the goodness of fit model (Bird *et al.,* 2014). According to Royle in 2013, linear and additive models assume that the response variation will follow a normal distribution, which is only suited a specific kind of measurement but may not be adapted for count or binary data which are described with other distribution (i.e. Negative regression or Poisson for count data and logistic regression for binary). Zero-inflated models are useful if a dataset contains a significant amount of null counts which can violate the Poisson or negative binomial distributions' assumptions (Arab *et al.,* 2008). In the cases where the closely spaced samples have more chance to present similarities with one another than with the one that are more spaced, an autoregressive regression model can be used (Pierre *et al.,* 2002).

The broad application and the relative accessibility of generalised linear and additive models still present limitations concerning the number of predictors and covariate they can describe simultaneously which demanded an extensive and crucial preliminary research on which model is the most adapted to analyses the data with the fewest parameters possible (Royle, 2013).

With a large amount of data, there can be an equally substantial number of predictors and covariates influencing the response variance, which machine learning can offer a more suitable solution. Besides, linear and additive models are generally, not the best fit for presence-only data except in the context of species distribution model and uneven variance across samples within the dataset (i.e. heteroscedastic) (Bird *et al.,* 2014).

4. Mixed-effect

Mixed effect models are a powerful tool to account for sampling bias; it can estimate the influence of predictors on the variability of the dataset without affecting the mean response by including fixed effects that are used in a linear or additive model with random effects (Royle, 2013). This is extremely useful in a situation where observers have different sampling efficiency as the mixed model will assume that if each volunteer contributed with one observation, the mean of these observations would then be centred on the real mean. However, it is necessary to attribute the observers with identification as if some of the volunteers contribute more to the dataset than others it can skew the overall average, and the identification could be used as an index to the model viability preliminary to estimate the influence of predictors (Schwarz, 2013).

Mixed models have proven to be useful in ecological studies due to their flexibility and predictive power; therefore, they also been able to accommodate observatory and spatial clustering bias (Bolker *et al.,* 2009).

5. Hierarchical models

When the sampling design of a crowdsourced dataset present element of systematic bias, a hierarchical model is a valid option to measure their influence. The hierarchical model presents similarity to the mixed models as they are used to describe the relationship between predictor and response variable but where the hierarchical model differs is that the parameter themselves can be described as a function of the other predictor variables (Schwarz, 2013). The model can be adapted to match specific citizen science projects and deal with false positive, misidentification, and detection imperfection; however, hierarchical models usually require specific sampling design to be able to correctly and accurately describe the process (Bird *et al.,* 2014).

V. Some words on new technologies and Artificial Intelligence in Citizen Science

The last decades have witnessed an unprecedented technologic expansion allowing to collect an enormous amount of data. The web 2.0 that started to emerge in the early 2000s permitted users to have a more powerful influence on the contents published by allowing them to directly post, edit, comment and provide different information on numerous platforms and networking tools (Galaz *et al.,* 2010).

There are currently 3.6 billion internet users, just shy under 50% of the human population (Meeker, 2018),. However, the access to information technology is still unequally distributed with developed countries having and adoption rate close to 80% whereas the least developed countries, only one in 5 citizens have access to the worldwide web (International Telecommunication Union, 2017). The internet adaption rate has seen its growth reduced to 7% per year in 2018 globally due to the saturation of developed market but still have a double-digit adoption growth in developing economy (Meeker, 2018).

Along with the rapid development of information and communication technology, with the increased flow of data, there has been an evolution on the uses of internet available information in several creative ways to take advantage to help environmental causes.

The advancement of technology helped scientist to go away from expensive devices (e.g. radio collars) to smaller and smarter (e.g. satellite tracked) devices and even innovative non-invasive approaches (i.e. photo recognition) (Pimm *et al.,* 2015).

One area where existing technology has seen a significant improvement on the quality of data collection is through live location-based mapping services by only using the volunteer´s smartphone automatic capture feature (Ko Lwin & Murayama, 2011). Mobile phones are becoming personal measurement tools and will offer yet-unimagined services, but the volume of data generated by those advancements will require proper management capabilities.

The increased use of web services helped to develop the automation of computer to computer interactions with metadata generation and interoperability between large databases (i.e. Wild Me project). The challenge remains still acute, as for conservation biology species of interest are usually rare and count far fewer observations which can be drowned by the massive flow of data coming from millions of data entries (Pimm *et al.,* 2015). Moreover, it sometimes needs human intervention to manually sort through thousands of pictures as most of the database to have exploitable results (Bowser *et al.,* 2013). Computer vision tools are started to be integrated into the crowdsourcing database to reduce workflow and add efficiency (**Figure 2**).

Operating within the broader field of artificial intelligence, computer vision is part of a category of deep learning algorithms which are a set of methods that make them learn from a dataset without human guidance (Di Minin *et al.,* 2018). The algorithm will work with multiple levels of representations that are obtained by transforming the representation at each level, starting with the raw inputs, into a higher layer of representation while discriminating irrelevant variation to allow a classification (Lecun *et al.,* 2015). By applying the classification that offers deep-learning algorithms to social media, they could analyse metadata on an unprecedented scale to address biodiversity crisis, as many social media platforms offer the user to generate texts, images and video along with other exploitable data such as time and location.

Processing the data by hand will be extremely time consuming and excruciating, but the machine learning algorithms can be trained to distinguish, and filter specific aspect of the content analysed. They can detect specific species, determine if the picture is real and even the individual if patterns are visible (Di Minin *et al.,* 2018).

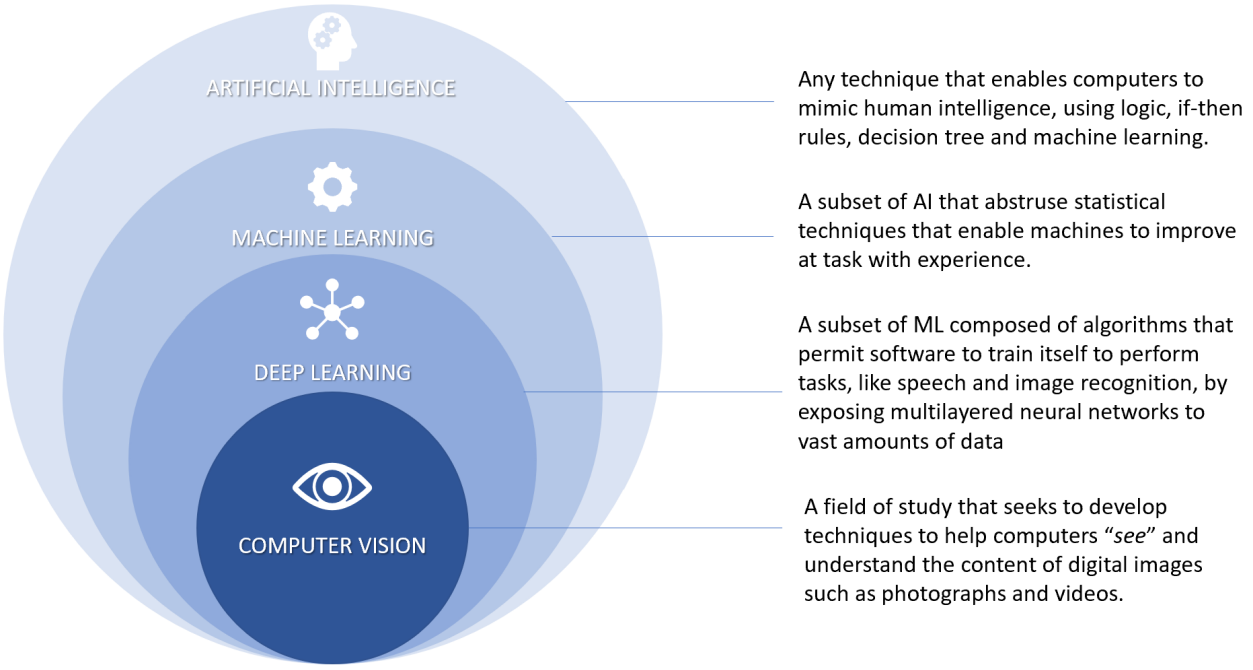

Figure 2: Diagram displaying the differences and relationships of the fields contained in artificial intelligence technology

For this review, we focused our approach on the implications that the deep-learning algorithm, coupled with the usage of social media in citizen science. By exploring the feasibility of this technology by creating a simple neural network and test how the algorithm would react to the different pictures and comparing the output results. Analysing and exploiting metadata to give a proficient assessment of the quality of the algorithm depending on the input pictures so we can learn what the limits of this process are.

What this thesis is all about: The ***Wild in Live*** **project**

The ***Wild in Live*** project was created for the development of a simple task: Is it possible to save time and effort, by extracting valuable data directly available on a picture of sea turtle and be able to match that individual with a database containing thousands of other sea turtles? Image recognition is a critical aid for wildlife conservationists and is used on many studies requiring data on habitat use, distribution, behaviour and population (Van Tienhoven *et al.,* 2007).

By recognising patterns and other natural marks on an animal, the photo-identification technique is a non-invasive tool ideal to identify a specific individual which is threatened or endangered (Reisser *et al.,* 2008). The wildlife photo identification technic pioneered by Dr Bigg in 1976 (Obee & Ellis, 1992) is not by any means new, but the improvement of the technology have allowed scientists to process far more data than a single ever human could. With the introduction of machine learning algorithms in ecological studies, the potential for these advancements will be even more significant (Kotsiantis, 2007). By Pairing the ability of scientists to social media data, it could bring another set of tools that will be able to interpret and visualise the vast and ecologically untapped data available on social media platforms. With the aid of machine learning, an algorithm can be specifically designed for image recognition could be created and then be used to identify a specific species and recognise its features that allow for individual recognition.

The metadata collected along with the picture on social media (e.g. location, date) are also exploited to give more information on the whereabouts of the individual identified. The automated collection and analysis of data will act as a passive citizen science programme where the inputs are made passively by the users of social media platforms and actively exploitable by conservation scientists. All while reducing bias that humans tend to fall into while collecting, sampling and monitoring during citizen science project.

I. Aim and Development steps

For this project to be fulfilled, different goals have been set to ensure the creation of a tool that is serviceable for conservation by scientists and amateurs alike.

1. A database, to save all the collected information with the algorithm with futureproofing to be available for other databases.
2. An identification tool algorithm capable of automatically roam social media platforms collecting and analysing pictures.
3. An intuitive display of the data and its results with professional features for customisation.
4. The ability to download and upload data directly on the database and be shared freely with other conservation organisations.

By making the project accessible to scientists and citizen scientists, the impact on the data collection and confidence could be increased as the immense amount of data collected.

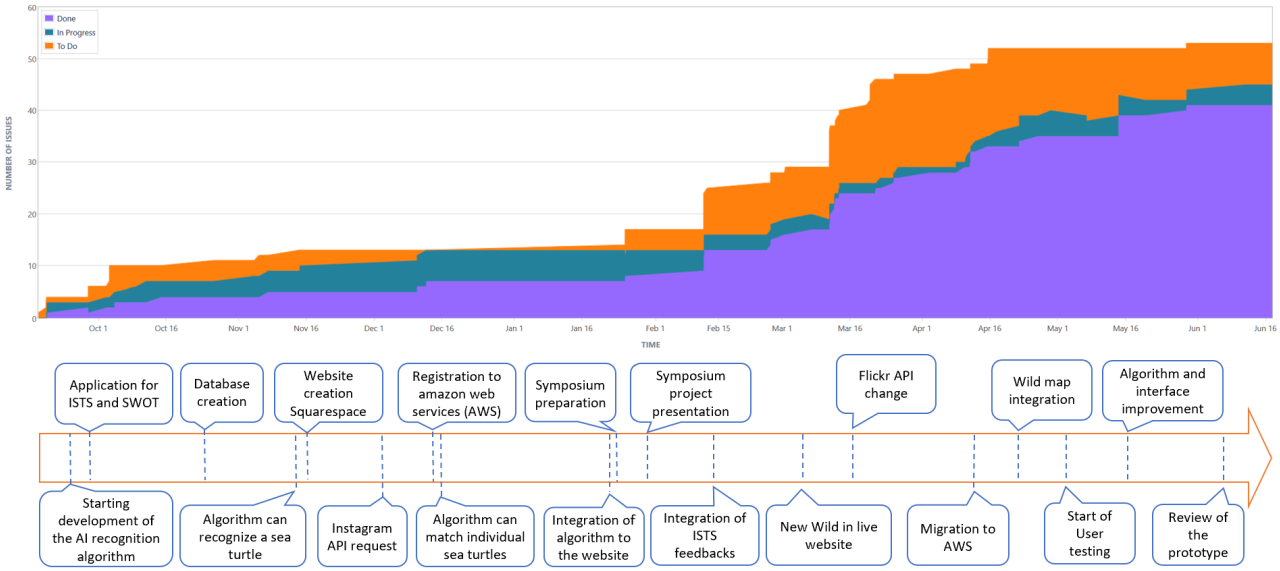

Figure 3: Development diagram of the *Wild in Live* prototype at Critical; it represents the number of issues and tasks completed overtime with a chronological timeline of the main event that occurred during the project.

The ***Wild in Live*** project started in September 2018, but the actual development was accomplished during the last six months, the project is created in collaboration with the Fikalab (which is the incubator from Critical Software) that provided the necessary resources for the development of the project.

There were two crucial deadlines for the development of the prototype; the first one occurred two months into development when the first version of the prototype was presented during the international sea turtle symposium in Charleston, South-Carolina.

This event was the occasion to get preliminary feedback on the project directly from the scientific community working directly with sea turtle conservation and tailor the project by recovering useful insights in the matter before starting further development. The second deadline, at the end of the six months of development, was a review by the Critical software of the state of the project and to discuss further development (**Figure 3**).

II. Description of the operation of the algorithm

1. Architecture

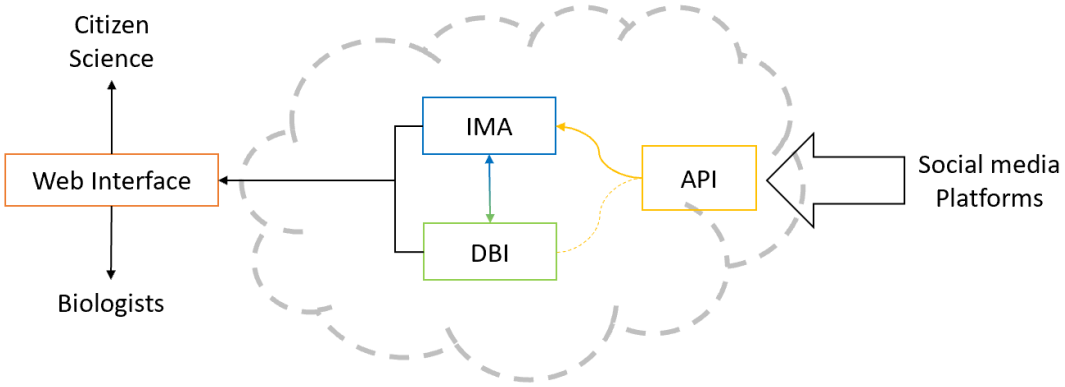

Figure 4: Schematic of the architecture of the *Wild in Live's* processes

The structure of ***Wild in live*** is shown in **Figure 4**; the system is composed of an Image Processing Algorithm (IMA), a Database and Interaction (DBI) and a data retriever (Application Programming Interface or API). The interface includes options to access identification tools and monitoring results computed on servers always available online as a web application running on Flask © (version 1.02) with PostgreSQL © (version 11.2) managing the data. Since the project is still a prototype, different clients were available depending on the web-based service available throughout the development (Amazon Cloud Service is our final one).

The IMA contains image pre-processing tools to improve the picture data by suppressing undesired distensions and by enhancing some of the features for later processing and analysis. The DBI and IMA interact closely together to implement a workflow that each picture undergoes to be usable data imported with the API.

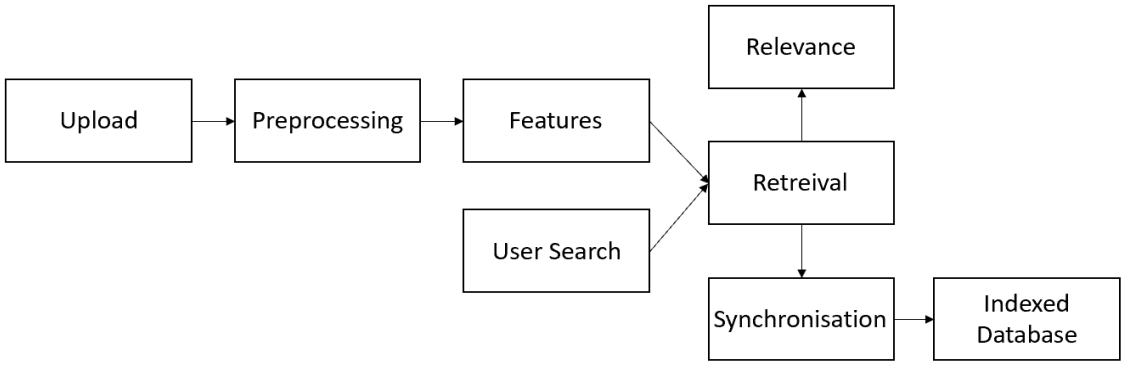

Figure 5: *Wild in Live's* workflow is an interaction between the users and the system and includes preprocessing, feature extraction, user search retrieval and relevance feedback. The information is then synchronised n the indexed database.

The workflow depicted in **Figure 5**, typically involves two different possibilities with a passive data collection and active user search in the database. The first option results from salvaging images and metadata from social media (i.e. Flickr ©) through the API script. Once the images are pre-processed for correct illumination and geometry, the features are extracted to be ranked and compared to the data available in the system, a relevance score is produced and stored. The second option let users to directly look through the database using natural language to retrieve information collected by the API or upload its data to be processed and analysed.

The output of the system is a table with all the sighting and identification information of the researched individual or if a result came back without matching, the ability to add or edit entries in the database if the user had uploaded the picture. The DBI and IMA can work asynchronously and operate in different modes depending on the user's interaction with the web application (e.g. identification, monitoring, data upload /download). ***Wild in Live*** enables users to work on a different aspect of the workflow in parallel, since some users may help the identification process by giving feedbacks if a picture is misidentified with another by adding sighting data on an identified individual. In this case, the DBI register the changes in the cloud, and the IMA improve its detection and matching algorithm.

It is necessary to implement in such a system state a harmonisation across the multiple steps to reduce different inputs made by users or algorithm. Therefore, all the picture that shared the same identity will be merged or created as needed, with the option for the user to update the information available.

2. *Why sea turtles?*

The ***Wild in Live*** project is focusing on sea turtles for proving its concept as they are the perfect candidates to start with. Sea turtles are under threats from rising sea water which is a loss of habitat for the nesting beach (Fish *et al.,* 2005), fisheries bycatch (Lewison & Crowder, 2007), nutrient availability (Chaloupka *et al.,* 2008), poaching and illegal trade (Mancini & Koch, 2009) that are the results of human impacts making some population of sea turtles endangered (**Figure 6**).

Figure 6: Poaching activity, credit to primeirasesegundas

However, sea turtles have an impact in the collective mind especially in western culture as it is an ambassador for the marine environment, a charismatic representant of its habitat and of the impacts that it is being threatened with. Sea turtles are easily recognisable by almost everyone and can be used by conservation projects to appeal to their causes more effectively.

They are by definition a flagship species. Being easily recognisable is crucial as it allows for a more efficient identification by social media users that may recognise a sea turtle and add additional metadata. Being a relatively large vertebrate will allow having an overall better-quality picture as more features of the animal can be identified (e.g. scale patterns) with the limited resolution imposed on most of the social media platform (between 320 x 320 and 1080 x 1080 pixels for Instagram). The scale patterns are used as an identification feature to recognise an individual (Reisser *et al.,* 2008) valid for the entire life cycle of a sea turtle (estimated to be around 80 years). Having a lifespan longer than most species, research projects will increase the chance of sighting of the same individual by a random encounter with users of social media poster, thus increasing the chance of collecting more data passively for the detection algorithm.

III. Development of the identification and classification algorithm

1. Current methods

In order to define how the technology used in this project works, an explanation of the current level of software devolvement is required. Most of the models available are coded in a spreadsheet or a monolithic bloc code in various programming languages that are developed as an in-house project for unique cases. With often a documentation not matching the complexity of the displayed code which makes the code difficult to be reused and adapted for other projects that often need to start from the beginning of development (Pauliuk *et al.,* 2015).

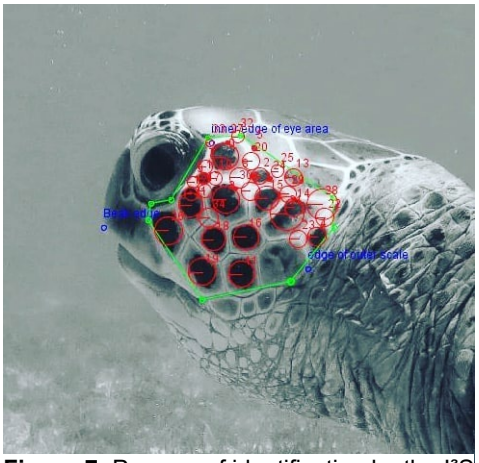

Figure 7: Process of identification by the I³S software; each scale is a reference point that will be analysed after affine transformation as coordinate.

Currently, one of the most used identification software was developed by Van Tienhoven *et al.,* in 2007, called the Interactive Individual Identification System (I³S). The software will use a two-dimensional transformation to compare two individuals in a similar setting and defined space. It requires the user to manually map the features of the animal studied (i.e. spots or scales pattern), then the software will convert the picture into two dimensions by the affine transformation which enables for the mapping of the marking and feature present on the individual onto a matrix of a coordinate **Figure 7**.

Then the comparison is achieved by comparing the source points onto the destination points, and the best score from all the possible transformations is used as the final matching score between the compared images. The software can have a correct chance of identifying the same individual of 95% if more reference images are present in the database (Van Tienhoven *et al.,* 2007).

The main feature of the I³S software is that human intervention is still used to judge and distinguish between the features and artefacts (e.g. reflections, shadows and particles) present in the pictures.

If the I³S software offers scientifically exploitable results with high confidence, it relies on the need for human intervention to judge each individual picture, implying that the researcher conducting this method, would spend hours to confirm an identification of multiple individuals. Moreover, it would require to manually enter additional information collected by citizen scientists such as location and date or distinguishing features as the software does not offer to group those under one bundle. This constitutes a substantial limitation when it comes to processing a large amount of information.

Unfortunately, I³S have not received any updates since 2014, and as discussed before, those software developments are dependent on the creators and institution managing I³S. Since the code is open, there is a real possibility that another researcher is using a more recent version, but without any distribution, the current version of the software lacks any advancement.

Therefore, by ensuring that the ***Wild in Live*** project is as open as it can be and available online, newer features will automatically be available and distributed to all and the uniform code (Python) could be adapted for other research and monitoring projects. Python language allows programmers to focus more on the design than the coding, easing the creation of a prototype and experimenting features that could then be easily implemented. Python is also one of the most popular languages when it comes to machine learning, as the code is free, open source and commonly used most of the computer-vision cases are already available in Python libraries for scientists (Pedregosa *et al.,* 2011).

The picture, along with its metadata (date of creation, GPS coordinate) and tags (location, input name from the user) are retrieved directly from the Flickr servers. The data is made available through Flickr's system application programming interface (API). The Flickr API enables third-party services to communicate directly with Flickr's servers and access the users' data with special authorisation (Nov *et al.,* 2008).

2. The deep learning technology

As mentioned earlier, deep-learning algorithms offer an adapted solution for automatization of image recognition with supervised learning. These algorithms need human intervention to provide input during the learning mechanism, which in return get feedback on the task performed for the creation of the model or the classification function (Thomaz & Breazeal, 2006).

a. Training of the machine

The training of the machine happens on Tensor Flow© which is a development platform that enables the facilitation of a large number of computational resources for training models on large datasets and moving them into production (Xu *et al.,* 2016). The learning starts by "feeding" the algorithm with the collected data set of 22,500 pictures representing each of the five previously defined categories. The dataset is partitioned for training and validation purposes, using 75% and 25% of the images of the training dataset respectively.

For the ***Wild in Live*** project, the classifying machine is shown an image, and it then produces an output in the form of vectors of scores, one for each category. Each category is assimilated to features to help determine what the picture is representing to be then able to differentiate between species of sea turtles or if the picture is showing a juvenile or a dead individual while getting more information out of the data (**Figure 8**).

The categories chosen for differentiation are:

A. Sea Turtle
B. Juvenile Sea Turtle
C. Sea Turtle Laying Eggs
D. Edited Sea Turtle
E. Tortoise
F. Not Sea Turtle

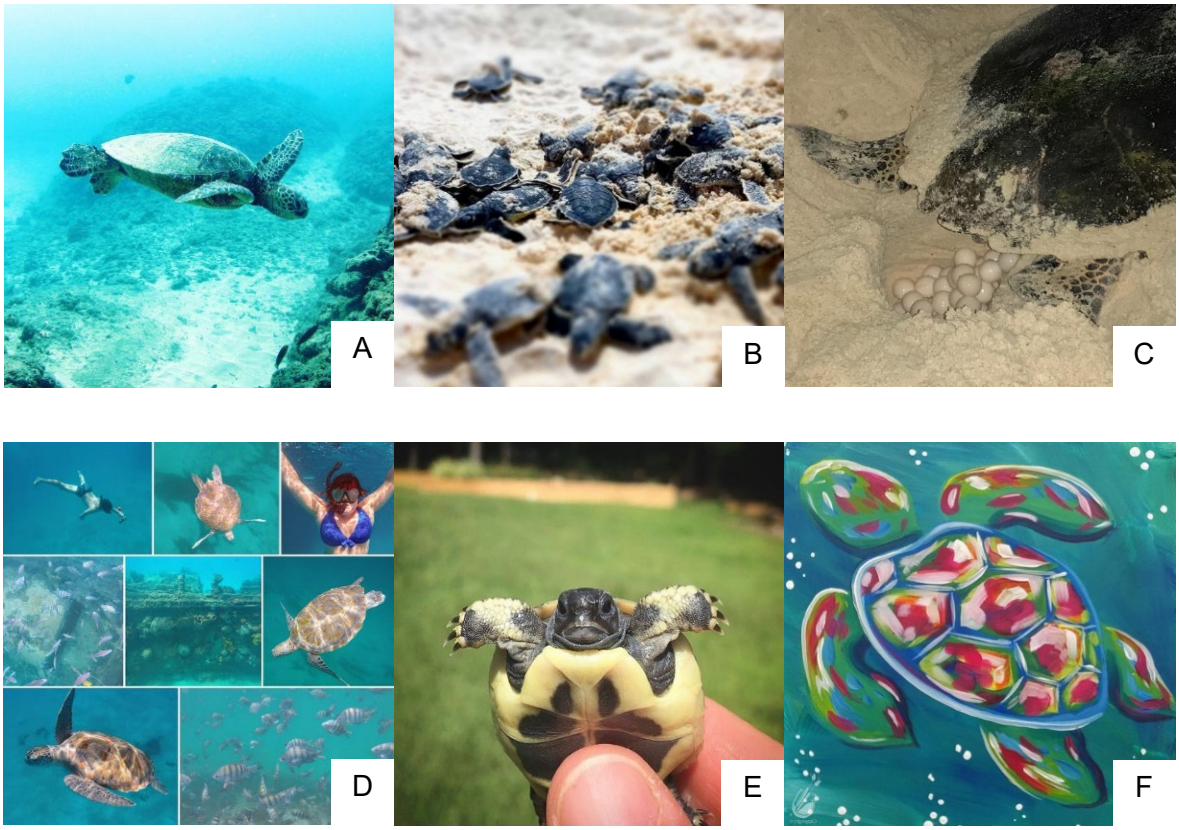

Figure 8: Pictures of each different categories that can be processed by the algorithm.

During the training, the desired category (sea turtle) should get the highest vector score of all the available category. The errors of the categories are computed by an objective function that measured the difference between the outputs scores and the desired pattern of scores. The machine will then automatically adjust its internal parameters to reduce this error. The parameters are often referred to as “knob” and define the input and output function of the machine. Typically, a machine has millions of these adjustable weights along with hundreds of millions of labelled examples to which the machine can train. For the proper adjustment of the weight vector, the algorithm calculates what would happen to the error if the weight was changed by a minuscule increment either increased or decreased.

The process is repeated until the average of the objective function stops decreasing. Then a performance review of the system is measured by testing the machine on a different set of images to determine its ability to produce sensible answers on inputs it has never encountered before. This entire training process is referred to as an “epoch” (or generation), and a machine usually goes through hundreds of them before offering satisfying results (Lecun *et al.,* 2015).

b. Image recognition

The purpose is to identify which features that are extracted from a picture can allow being correctly matched with a high probability against an extensive database containing the features for sea turtle image recognition. To extract those features without looking each pixel of the image, which would need exponentially more computing powers. The images are processed through a cascade of filtering approach in which the more demanding operations are only applied in the location of an image that passes an initial test.

Wild in Live's machine uses computer vision and therefore focus its computing power on the interpretation of features from the image that are generated through four different "layers".

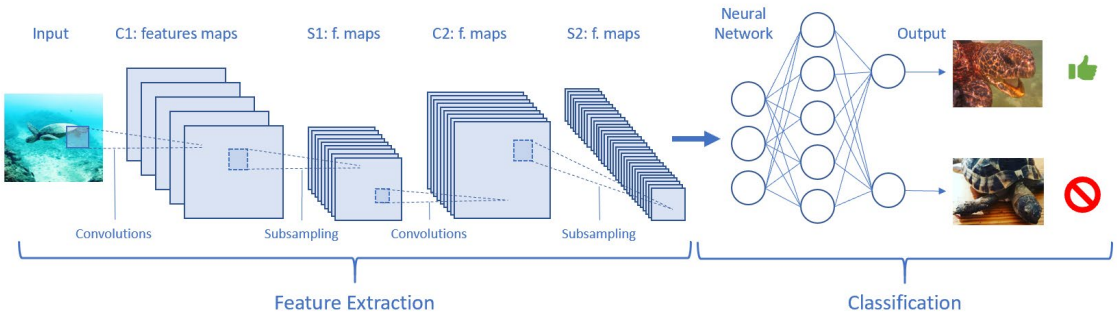

Figure 9: Example of a convolutional layer network.

These layers can interpret input by transforming it by performing a convolutional operation and create a modified output that is then picked up into the next layer. With image recognition, each convolutional layer has a specific number of trainable filters, which are applied to the input image and will allow the detection of patterns (**Figure 9**). The first layer of representation will look for the presence or absence of edges at a singular orientation or location within the picture. This layer is defined as the Scale-space extrema detection and uses the differences of a Gaussian function to identify potential interest points that are invariant to the scale and orientation of the image. The second layer will look within each candidate location detected in the previous layer to determine key points which are based on their stability from the localisation and the scale. The third layer will add one or more orientations to each key point based on the local image gradient directions. Therefore, all future operations performed on the original input image will only require the assigned orientation, scale and location for each feature and thereby providing invariance to these transformations. Finally, the fourth layer will measure the gradient of the image at a determined scale in the region around each key point.

The key point descriptor layer can assemble motifs into a more significant combination to correspond to part of a familiar object, and therefore, the layer would detect an object as amalgamations of those previous part (**Figure 10**).

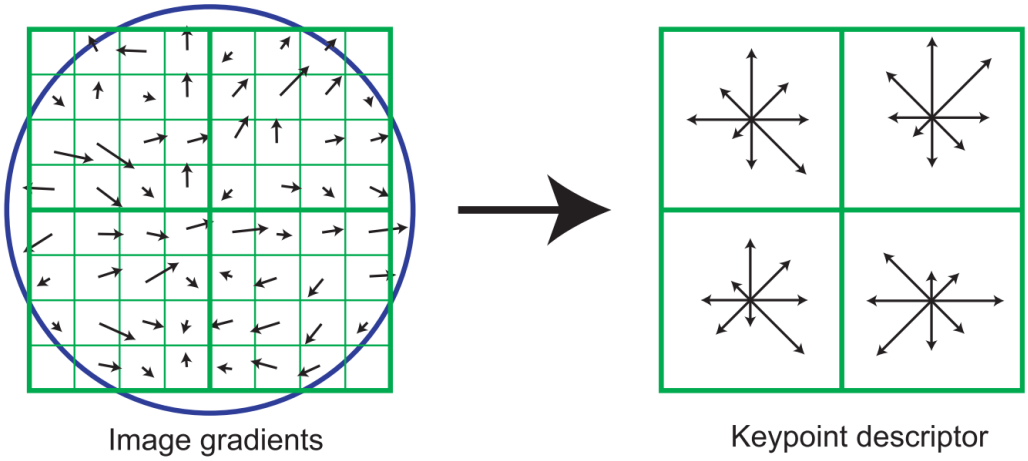

Figure 10: A keypoint descriptor is created by first computing the orientation at each image sample point in a region around the keypoint location, as shown on the left. These samples are then accumulated into orientation histograms summarising the contents of the region, as shown on the right, with the length of each arrow corresponding to the sum of the orientation near that direction within the region (David G. Lowe, 2004).

These layers create a structure defined as the convolutional layer network (CNN), which is the most widespread model in machine learning technology for image recognition (Wang *et al.,* 2018). The result of the image transformation through the different layer is a 2-dimensional activation map with scale-invariant coordinates relative to local features.

This approach has been named by David G. Lowe in 2004 as the Scale Invariant Feature Transform (SIFT) and allows for a typical image of 500x500 pixels to deliver about 2000 exploitable features. A higher quantity of discernable features in an image is extremely important as it allows to increase the probability of a reliable detection by the algorithm even if an image recognition requires at least three features to create the allow the matching ability.

Only the key points and key points descriptors of the features linked to an image saved in the database. It allows to optimise the algorithm's performance as it only needs to look through numerical values of the 2-dimensional activation map to be able to identify a specific individual. A new image is matched by individually comparing each key point from the new image to the database and finding candidate matching key points based on the shortest distance between two points (also named the Euclidean distance) of their feature vectors.

IV. Application to the Wild in Live prototype

1. Turtle image classifier

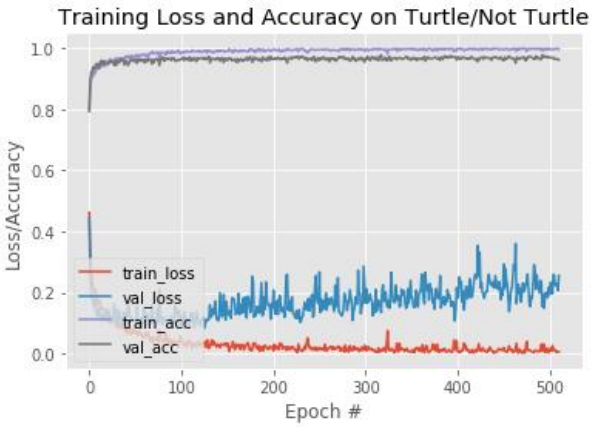

Figure 11: Graph demonstrating the accuracy and loss from training and validation

Several configurations of the algorithm were tested to obtain the best accuracy with minimal validation and training loss.

Those values allow to evaluate the performance of a supervised learning algorithm, the loss function (also referred to as “cost”) measures the divergence between the predicted output and the desired output value.

The best algorithm iteration happened with the 287 epochs, with the initial learning rate equal to 1e-3 and had an accuracy of 95.53% and presented lower validation and training loss.

However, as seen in **Figure 11**, the validation loss while still higher than training loss is increasing toward the 400 epochs, which means the model is starting to overfitting. Which happen when the model starts learning patterns from the training data that does not generalise to the test data and need more training images to improve. To reach accuracy close to 100%, calculation estimate that half a million of training picture is required in each category; totalling 2,5 million pictures to develop an algorithm with near perfect recognition.

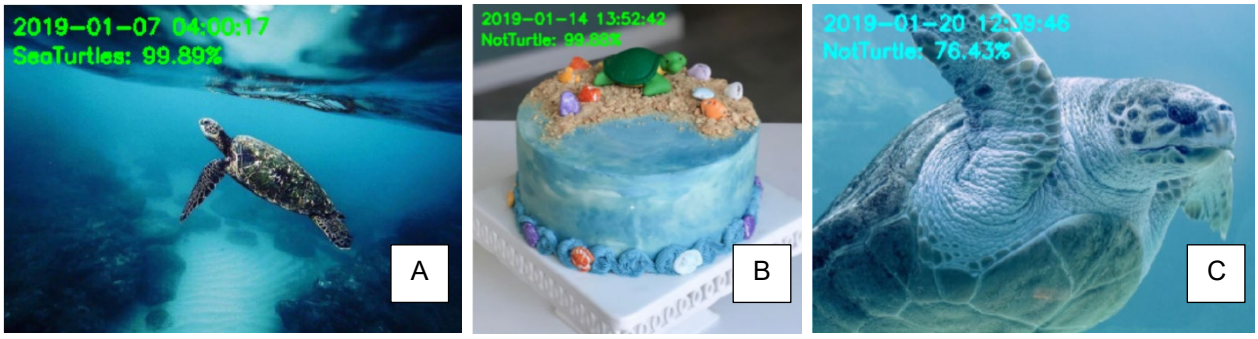

Figure 12: Result of categorisation by the algorithm. Picture A and B show a correct interpretation of the subject in the picture, whereas picture C is a false negative due to the angle of which the picture was taken.

Testing of the classification algorithm shown to work extensively well within a delimited range. If the picture is displaying a sea turtle in its entirety (full body picture), the algorithm can classify the input with a variance of 9.52% (**Table 1**). However, due to limited training pictures for tortoise, the algorithm will struggle to differentiate between sea turtle and tortoise and can misclassify a sea turtle due to overfitting (**Figure 12**).

2. *Matching turtles*

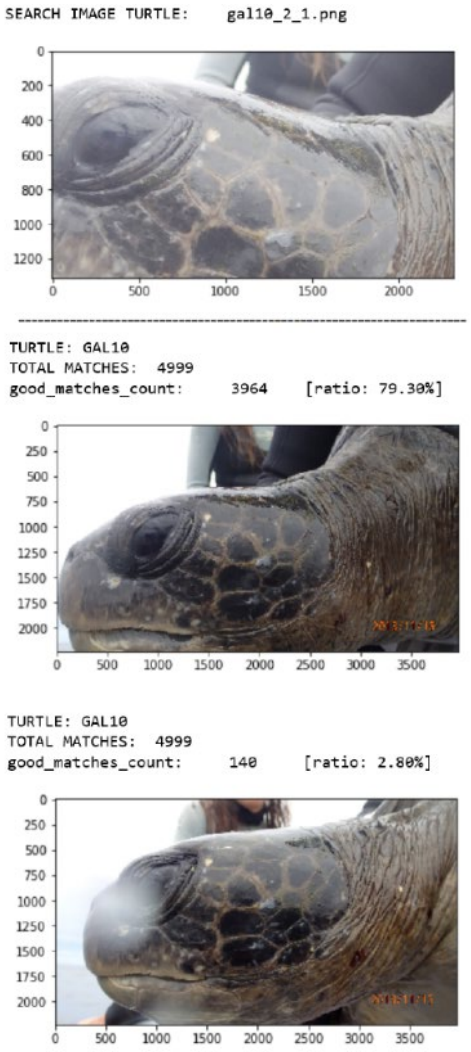

Figure 13: Individual identification process. The higher the good match count, the higher the probability that this particular individual is the same as the two other pictures.

Once the picture has been categorised as a sea turtle, the identification process can start to extract key points and descriptor from the input picture. The extracted data points are then loaded to the database and compared to other pictures' data points. The best match is the pictures having the highest number of similarities between the data points and the analysed picture, called "good matches" and consider the ratio of the closest distance to the second closest distance. The results are then ranked depending on the matches points and displayed to the user **Figure 13**.

The combination of the two processes of identification and matching make possible the automatization of the monitoring of sea turtle across social media platforms. They both need to perform appropriately in their task to increase the reliability of the process.

The performance for each category depends on the difficulty of the task presented and the number of training pictures available. As the "Sea Turtle" and the "Not Sea Turtle" have 0.163 and 0.168 average times for categorisation and with the lowest variance possible in their confidence score, they can carry out the full identification in a reduced amount of time (**Figure 14**). However, categories with training picture will have increased variability in their confidence score as the algorithm is not trained on enough data to be operational as a scientific identification tool.

This effect can be illustrated by both the “tortoise” category and the “juvenile sea turtle”. They have both lower average in their confidence score; the ‘tortoise” one is due to the resemblance to the better established “sea turtle” category, which increases the identification time. Whereas juveniles are harder to distinguish from small artefacts as their feature are smaller than most of the other categories making the categorisation and identification more dependent on the quality of the input picture (**table 1**).

Table 1: Reliability of each category with the current iteration of the algorithm

Identified Class	Confidence Score in %	±
Tortoise	82,58	13,88
Not Sea Turtle	93,61	11,45
Sea Turtle	95,53	9,52
Juvenile Sea Turtle	81,92	20,68
Sea Turtle Laying Eggs	93,50	16,76
Edited Sea Turtle	84,63	15,42

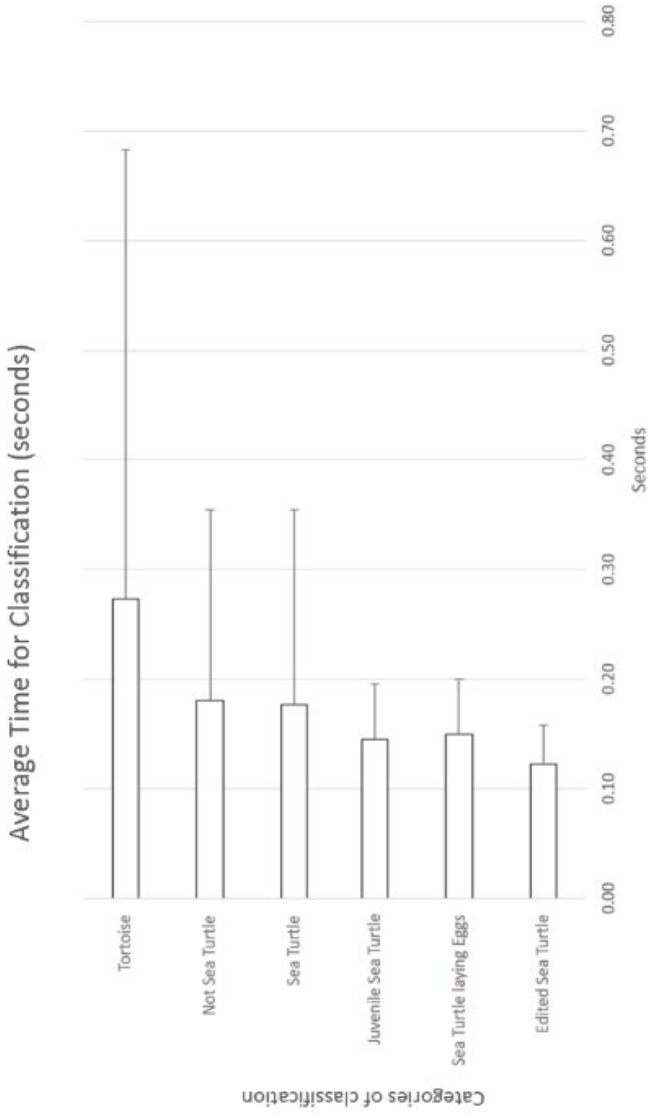

Figure 14: Performance of the Wild in Live algorithm for each category

3. *The interface: the Web application*

The ***Wild in Live*** prototype is available online (https://wildinlive.org/) and offers different features to collect and sort data on social media while allowing the consultation of its database (**Figure 15**). However, some of the features are still in development and not fully completed as more testing are being conducted.

1. Social media search: the feature will use the API to communicate directly with Flickr´s server and analyse the last pictures posted (capped at 30 in the wait of better optimisation) and detected if a sea turtle is present (**Figure 16**). If so, the picture is displayed for further analyse, such as individual recognition (**Figure 17**).
2. Identifying individuals: if a user owns pictures of sea turtles, this feature enables to upload the sea turtle picture to the prototype directly. Then the recognition algorithm will analyse then to find potential matches already present in the database and display any additional information available on the previous sightings of the individual (**Figure 18**).
3. Import data: if a researcher has sighting data along with pictures, it can be imported directly in the database so the information can then be treated and made available for later search (**Figure 19**).
4. The Live Map: is the most ambitious feature in development. Its purpose is to display the data from the database in a different way depending on the natural language input of the user. The user could look for the specific data source in a search request like what a user could ask on Google, and only select what is significant all with the ability to download the data (**Figure 20**).

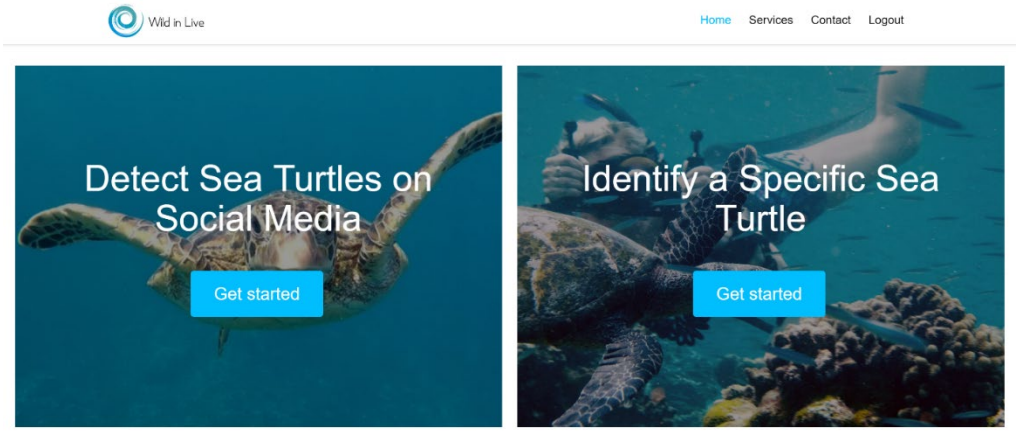

Figure 15: Screenshot of the home page from www.wildinlive.org, the two main option for citizen science project “detection” and “categorisation”. Other services are available in the menu bar but are only available once the user has logged its credentials.

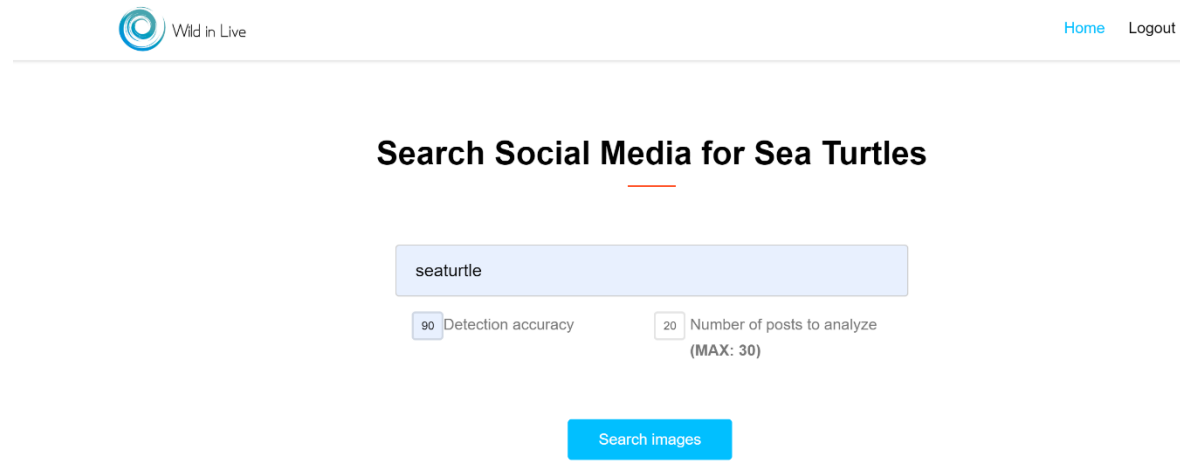

Figure 16: Screenshot of the detection page. It allows to look for hashtags or title through social media, and the algorithm will fetch the results in the last pictures posted with an accuracy previously determined.

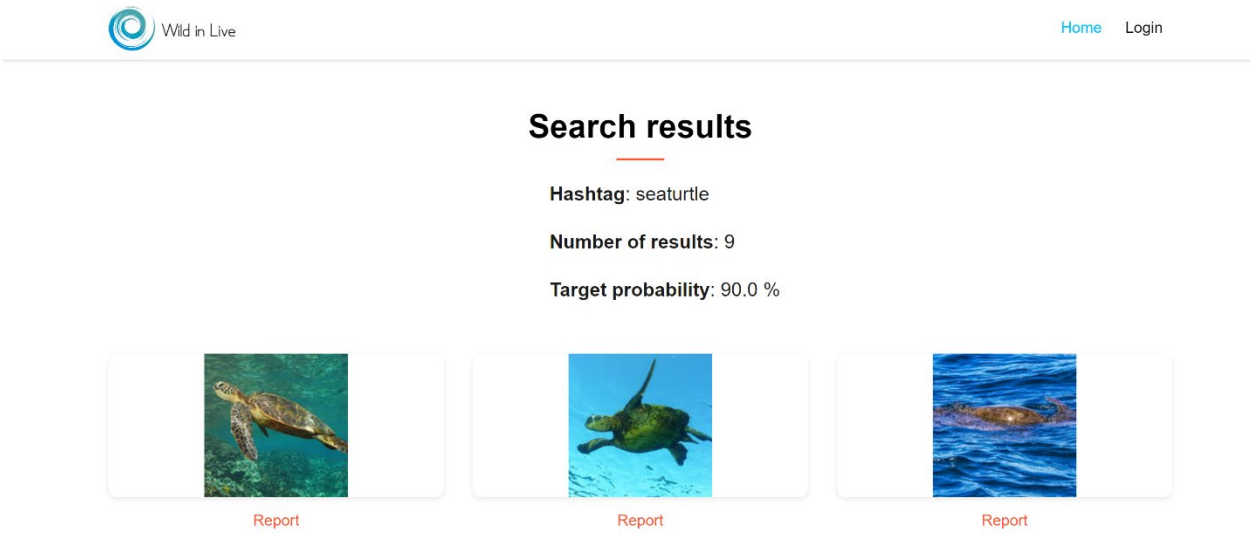

Figure 17: Screenshot of the detection result page, each picture is clickable and with the option for individual identification. If a picture is not a sea turtle, the user has the option to report the result as feedback to the algorithm.

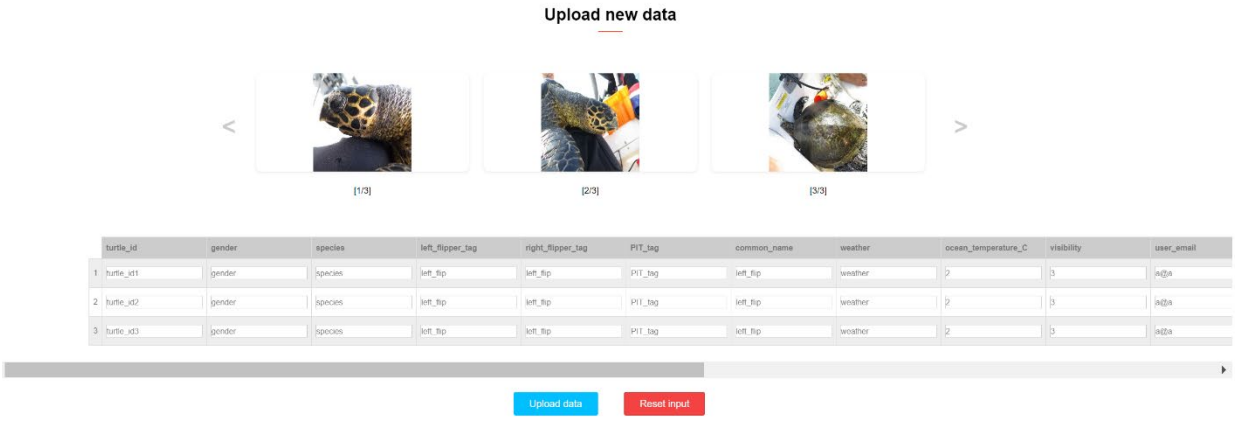

Figure 18: Screenshot of the identify turtle option. A researcher can import its pictures directly into the website to be analysed and possibly matched with the database.

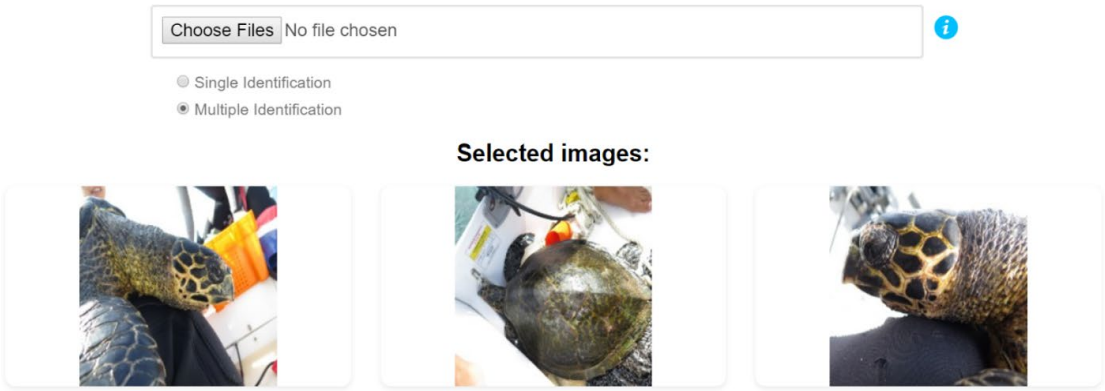

Figure 19: Screenshot of the import data option. A user has the option if none of some of its pictures have not matched in the database to upload them along with the sighting data directly into the dataset

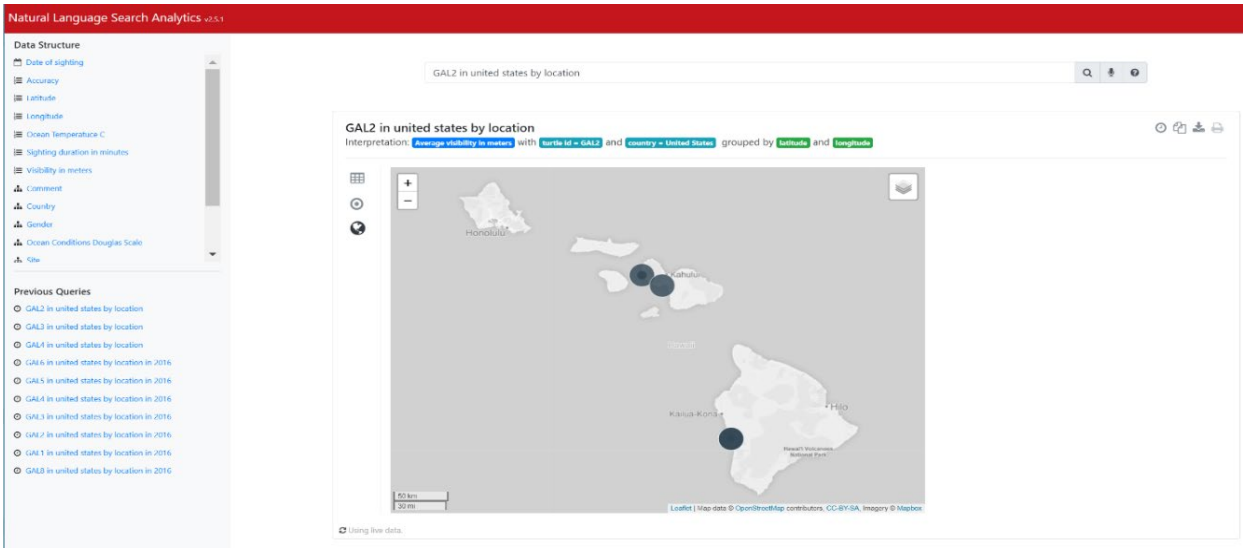

Figure 20: Screenshot of the Wild map. Here the user can search the database for specific information that is available. It allows to have all the data easily accessible and d downloadable for further analysis

V. Issues encountered during the development of the project

1. The training of the algorithms starts with manual inputs.

As time-saving automation processes created with the help of machine learning are, it still requires a significant time to set up and especially for computer-vision processes. As seen earlier, to make an algorithm learn, it must be “fed” massive quantity of data to learn what to “look” for. However, to ensure that the data supplied to the algorithm is representing one of the categories in need of description, the pictures need to be *manually* sorted, checked and placed in files to be used either for training or validation. More than 22,500 pictures were collected through many conservationists willing to share their data freely to help the development of this project.

To complete the project and have a reliable classification algorithm, even more pictures will be needed, and a modified classification system inspired on ***Wild in Live***’s algorithm would be needed to sort through the bulk of the pictures with human supervision to guaranty the quality of the data. This process will be necessary before the ***Wild in Live***’s algorithm has reached a high scientific standard. Reaching this goal will require even more data that would need to be stored and handled correctly to reduce the risk of loss or misappropriation. If the ***Wild in Live*** project can be relatively simple to manage as of now, reaching this goal will demand an upscaling in mean and human resources to conserve high performance throughout exploitation.

There is a direct correlation between an increasing amount of data processed and the computation power needed to complete the task. With the current iteration of the algorithm, it processes an input query by comparing with all the pictures of the database, the algorithm risk to have its processing time dependent of the amount of data in the database. Different techniques can mitigate this issue such an orthogonal approach to partition data that avoid running the algorithm on an extended dataset (Kotsiantis, 2007) but are not implemented yet.

2. *Difficulty in distinguishing different sea turtles' image.*

Social media platforms and their never-ending stream of content often came with a great variety of closely similar pictures. Determining what a sea turtle is could seem a simple question. However, thanks to the advance in technology and computer-assisted editing software, their frontier between what is real and what is computer generated or modified makes it more ambiguous to identify as an authentic sea turtle. This issue is particularly relevant when collecting data directly from those social media platforms, as the probability of having a false positive registered in the database could in term reduce the validity of the data and deteriorate the capacity of the algorithm to identify its subject correctly.

Figure 21: Examples of pictures that are photorealistic of a sea turtle but could be misinterpreted by the categorisation algorithm

Despite the advantages of image recognition such as reduced manipulation of the animal, increased efficiency in recording and storing data without the problem of damaged tag, photo ID is not always error proof either. Although in recent year the advent of technology made the quality of picture crispier at each new iteration, poor quality photographs decrease the match success by increasing artefact of the picture to be misinterpreted by the identification algorithm (**Figure 21**) (Reisser *et al.,* 2008).

A filter must be placed within the categorisation process to sort picture coming from social media platforms. Not all the same pictures will have the same value in the "eyes" of the algorithm since some of them can be duplicate of a famous original post or edited post.

All of which can be considered valid input but the metadata behind each picture will indicate completely unusable coordinates or date as some individuals will have been reposted hundred times. To reduce the contamination, the database and possible overfitting, any picture not reaching a high confidence score will not be used (**Figure 22**).

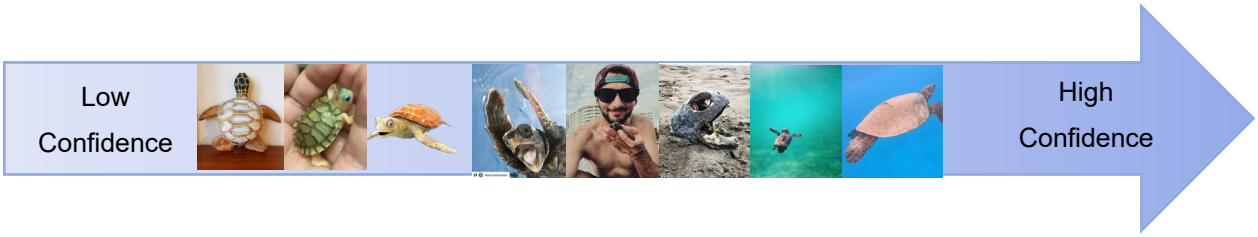

Figure 22: This diagram illustrates the variations when looking for sea turtles on social media. Not all picture carries the same information and, they need to be classified or filtered to ensure that the algorithm is only focussing on the "good picture" e.g. the pictures with high confidence of classification.

3. Authorisation of websites to get the data

The core of this project lies in its ability to fetch data directly on social media platforms, using API to communicate with the appropriate servers. However, since the Cambridge Analityca scandal that revelated in 2017 the misuse of data from Facebook by a third party to pollical purposed (Cadwalladr, 2018), the global outrage that rightfully resulted forced internet companies to revisit their approaches to personal data management. Subsequently tightening the rules to receive permission to tap in their resources, such as asking for a completely operational prototype before gaining access. This condition made accessing the Facebook-owned social media Instagram unavailable during the development of the project, setting a setback in the ability of the prototype to get real-time monitoring of sea turtles. As a default, the ***Wild in Live*** project had to default to the Flickr API as it offered less restriction in the admission request. However, it diminished the real-time monitoring feature since the picture posted on this platform are often edited before being posted, decreasing the spontaneity of the information.

Future developments and applications

1. *Gamification*

As talked earlier in the “Mitigating the challenges in Citizen Science” chapter (**page 9**), increasing the part of an active user is crucial for the viability of such platform as the input from scientist or wildlife enthusiasts contribute significantly to the quality of the data for a crowdsourced project.

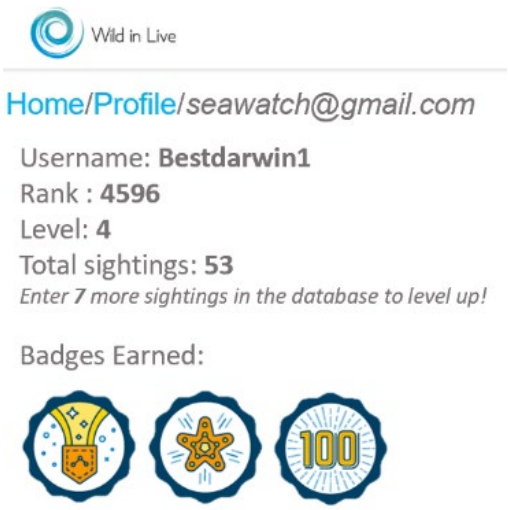

Figure 23: Mock-up of a user profile along with its badges and experience to motivate the user into being more active on the platform.

Moreover, one way to increase the active user proportion is through the gamification of such project. Gamification is the uses of game design elements in a non-game context, which can help create more enjoyment while interacting with an application. Citizen science is the right candidate for those kinds of opportunities as one of the significant volunteers’ motivation is an interest in technology and rewards (Bowser *et al.,* 2013). By integrating game-like elements such as leaderboard, badges and an occasional narrative to inspired volunteers, classification or reviewing tasks could be implemented at a lower cost of time and effort for the developers (**Figure 23**).

2. *Data background analysis*

One of the current limitations of the project is its ability to manage a large amount of data; as the prototype will retrieve data only when the user requests it. Therefore, the accumulation of data only occurs during the active use of the prototype, which can lead to missed collected information. To improve this aspect, the next version of the prototype should be able to retrieve automatically all the information posted on social media platform at any time to offer a real-time location of sea turtles around the world. It could allow the development of an alert system in the event of hatching or poaching as notification could be sent automatically to the concerned conservation organisations to deal with those issues as fast as possible.

Alternatively, if a researcher needs to know the whereabouts of an individual, a notification could be sent if the individual has been picked up in a newer location and time frame.

3. Estimate population and Migration patterns

Currently, the results available in the database mainly focus on the individual identification and will present factual data as it was collected, as individual point representing a sea turtle. These sighting data could also be used to offer additional information by including an estimate of the population depending in an area by including statistic model directly in the website, thus offering additional content to the user, and displaying the result in a more understandable way for later interpretations. The migration patterns could be created by linking the different sighting of an individual over time, displaying a network and heatmap of the whereabouts of the sea turtles.

The data collected with the monitoring could allow to calculate the average mortality rates, average recruitment rates, the size distribution of the population. Including those formulae in the algorithm could estimate the ability of a specific sea turtle population to persist in the future.

4. Alerts and notifications

A feature that was requested by sea turtle researchers is the ability to be notified if an event or sighting of a specific individual occurs. A researcher could then be informed if a hatching event is happening in the area under his supervision and send volunteers to protect and monitor the site. If a researcher is interested in a specific individual, ***Wild in Life*** could send a weekly review on its whereabouts. The alerts could also notify competent protection organisations if an illegal trade of sea turtle activity is detected on social media, as wildlife dealers often use social media to release photos and information about its product to attract potential customer (Di Minin *et al.,* 2018). The notifications could be easily be set up in the profile of the user but only available to researchers as it is sensitive information.

5. Ability to send feedback to improve the algorithm

As a supervised learning algorithm, human input is required to steer the system to identify its inputs accurately. However, as accurate the algorithm might be, some miss-categorisation or miss-identification might happen due to artefacts in the picture rendering the processing inadequate for the input image. Since the process is automatised, scanning a massive volume of pictures could create a significant number of unreliable results. A safeguard to this situation is through the gamification of the project, asking users directly to review the pictures themselves to correctly categorise them, increasing the certainty and reliability of the results while sending valuable information to the algorithm to learn from the feedbacks. If a picture cannot be identified by the algorithm nor by a user, it will be discarded to preserve the integrity of the database.

These implementations can be added with continuous updates on the project.

Critical overview of the project

The rise of citizen science has been proven indispensable as environmental preservation has become one of the main challenges facing our society. There is now a renaissance of citizen science through expanded government models, technological innovation and the increased interest in participatory and open science. The versatility of crowdsourced projects tends to provide a transdisciplinary paradigm, moving beyond isolated scientific disciplines and toward a matter of societal relevance across different fields. (Hecker *et al.,* 2018). It is with that very essence of innovation and interdisciplinarity that ***Wild in Live*** was created, to answer a simple problem using knowledge of both conservation biology and computer science.

Presently, the Wild in Life project is just a prototype and is not ready to be released in the hands of scientists or sea turtle aficionados. It is the evidence that real-time species identification through social media for conservation purposes is possible, but it still requires further developments to be able to provide the full set of services needed to be used as a scientific tool.

The development of the project shed light on matters arising from the creation of such an algorithm. One of the primary objectives was to reduce errors bias by reducing the human input in the data collection and to reach such level, the algorithm itself must be trained rigorously. Requiring that an automated process does at least as well as a human is not an easy task. A technological company like Google have been trying to render commercial self-driving vehicle since years and are still in the development process (Bimbraw, 2015). The stakes as not as high as ***Wild in Live*** is dealing in big data and not human lives and do not require an endless number of accreditations to get authorisation from a legal standpoint to be approved. However, even with a rudimentary training dataset, the algorithm can offer a sea turtle categorisation similar than human accuracy with a much-increased yield (Dodge & Karam, 2017).

Further finetuning of the algorithm will then help in the areas where it lacks performances due to less furnished training datasets, disordered data input and overfitting. The algorithm does not have the vocation of replacing professional researchers and would not be able to until advancement in computer-vison enables the algorithm to have more than a binary response to the input. The human visual system is better suited and has a more robust representation of visual stimuli than state of the art neural networks, being able to recognise subject under different conditions using past experiences and information available. An algorithm will only look for specific details in the picture, which might be unavailable due to the quality or the noise present in the image.

If presented with a picture containing ambiguities (i.e. picture with a dead juvenile in a nest), human's classification does not necessarily offer a more accurate classification than ***Wild in Live***'s, as each volunteer can have a different opinion on what the image represents. The advantages with a predictable process are that it allows for a better understanding of the outputs and we can easily identify where the ambiguities can lies, reducing the contamination of the crowdsourced database with low accuracy data (Russakovsky *et al.,* 2015). Besides, scaling up the dataset will always reveal unexpected challenges that will require a rewrite of parts of the algorithm's code and be often updated. Therefore, the growth of unlabelled dataset will imply to rely (until better iteration of the classification algorithm) on human supervision after the algorithm make its categorisation to verify its precision (estimates that the algorithm made and how many were deemed correct by humans). Once completed, the algorithm will minimise the effects of human-related bias in a crowdsourced project and be able to save time to researchers by scanning through thousands of pictures in one set.

To properly save time, performance must be noticeably higher than average volunteers carrying out the task. If the individual categories performance offers promising results with less than a second to get results, testing of the algorithm revelated that many other factors could make the experience significantly longer. The overall website performance lacks behind the standard of other applications users are accommodated to use and is the most case was either due to the local internet bandwidth or the algorithm was not correctly set up to handle an immense amount of inputs. These issues have been addressed multiple time during the development through optimisation of the database search and correction of architecture processes, but improvement is bottlenecked as only more power could improve the performances situation (i.e. more dedicated servers).

Here lies one of the main flaws of crowdfunding assisted algorithm, the user experience cannot be inadequate, or the project will be at risk of losing daily users and become irrelevant to scientists. To improve user experience, an intuitive interface and decent performance are mandatory to keep the project operational and therefore, higher investment is needed for the creation of the finished project.

The high initial investment is what makes the creation and development of these project delicate. Investing a significant amount of money in a project that might not be used by its niche user while maintaining update services and operational cost could explain why a substantial number of projects that are discontinued (e.g. Sloop, I³S).

Fortunately for researchers most of those projects have an open source directory available to enable customisation of the source code and update them to today standard, but it needs proper cyber-infrastructure investment in standardisation, interoperability, attribution, curation and preservation of the metadata to be used for crowdsourced project (Crowston *et al.,* 2012). The result is often the creation of a capable algorithm, but due to the lack of funding, it will only be used internally to the institution and not be easily shared with other projects, disseminating the scientific advancement in computer ecology to the only researchers skilled in this domain.

Wild in Live is using data available from social media platform and data uploaded by researchers, therefore, it must remain free and open to scientific projects as those conservation initiatives often run with a limited budget. Ensuring free access to scientific data need to come with a guarantee that it will not be used for non-conservation projects. Determine where and when sea turtle migrate, and their real-time location is valuable data that could be used as effectively to start poaching activity actively. Ensuring that the data is in good hands, will reduce the ease of access to information available in the database as not everyone should be able to download the database without any accreditation first. One way to deliver data to everyone while safeguarding sea turtles is to include in the algorithm a function that will approximate the data and will, for example, only disclose data to a country level of accuracy. For the raw data, a researcher will have to prove that its conservation project is indeed for the wellbeing of the individual and safeguards could be implemented to encrypt the raw data in the database in case of hacking.

Once the ***Wild in Live*** project will be fully functional and perform adequately, the algorithm will be able to identify to an exceptionally low degree of errors, that will come with an improved classification algorithm, bias inherent to the volunteer capacity to detect, identify and record the sea turtles. The immense amount of data collected over the years could help conservation community to have a more accurate of the state of biological threat that the seven species of sea turtles are facing; be notified when a hatching event occur and send volunteers to protect the area and be a great tool of communication to raise awareness. Working with this “big data” could enable the unearthing of patterns and the creation of new analytic models that were not could be developed before due to relatively low access to a massive amount of data.

Being an open data project (excluding Live map which is based on software developed internally by Critical), the code could be shared on other similar projects to help the spread of the computer vision technology at a significantly reduced cost to other conservation projects. Since development would already have been done, only a few modifications would be required to monitor other species.

Alternately, new species could be directly integrated to other iteration of the ***Wild in Live*** algorithm by enabling the researchers to import a large quantity of training data to create a new version of a computer vision process.

The ingenuity of this learning process is that the layers are not the result of human conception, but they are learned from using a general purpose learning process. (Lecun *et al.,* 2015). Exactly like it is impossible to explain how neurones in a human brain interact to identify and recognise an object, the CNN will create a network between neurone that will have the same ability without the ability to explain it how it developed it. However, with an essential advantage that it can be copied and improved with minimal resources which give machine learning an evolution capacity much higher than nature's (Burrell, 2016).

Machine learning tools have been used for more than two decades but only now with the advent of the internet of things that an incredible amount of information is available, that a new way of collecting conservation data is possible. ***Wild in Live*** proved that it could collect data automatically and if integrated to a citizen science project, it could become a tool that will enable an improve passive monitoring by decreasing the time needed to collect data while guarantying its reliability. Other challenges not foreseen will inevitably arise with the adoption of new methods to related to machine learning, but we are eagerly awaiting the future development of object recognition datasets and algorithms, as the possible application will surely have a positive impact on the way crowdsourced projects and conservation effort are conducted.

References

Aarhus Convention. (1998). *Convention on Access To Information , Public Participation in Decision-Making and Access To Justice in Environmental Matters*. *Aarus Convention*. https://doi.org/10.1017/CBO9780511494345.010

Andrew Hendry, C. P., Museum, R., Mimura, M., Yahara, T., Faith, D. P., Vázquez-Domínguez, E., Hendry, A. P. (2017). Understanding and monitoring the consequences of human impacts on intraspecific variation. *Evolutionary Applications*, *10*, 121–139. https://doi.org/10.1111/eva.12436

Aono, Y., & Kazui, K. (2008). Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century. *International Journal of Climatology*, *28*(7), 905–914. https://doi.org/10.1002/joc.1594

Arab, A., Wildhaber, M. L., Wikle, C. K., & Gentry, C. N. (2008). Zero-Inflated Modeling of Fish Catch per Unit Area Resulting from Multiple Gears: Application to Channel Catfish and Shovelnose Sturgeon in the Missouri River. *North American Journal of Fisheries Management*, *28*(4), 1044–1058. https://doi.org/10.1577/m06-250.1

Atkinson, I. A. E., & Cameron, E. K. (1993). Human influence on the terrestrial biota and biotic communities of New Zealand. *Trends in Ecology and Evolution*, *8*(12), 447–451. https://doi.org/10.1016/0169-5347(93)90008-D

Behmel, S., Damour, M., Ludwig, R., & Rodriguez, M. J. (2016). Water quality monitoring strategies — A review and future perspectives. *Science of the Total Environment*, *571*, 1312–1329. https://doi.org/10.1016/j.scitotenv.2016.06.235

Bela, G., Peltola, T., Young, J. C., Balázs, B., Arpin, I., Pataki, G., Bonn, A. (2016). Learning and the transformative potential of citizen science. *Conservation biology : the journal of the Society for Conservation Biology*, *30*(5), 990–999. https://doi.org/10.1111/cobi.12762

Bimbraw, K. (2015). Autonomous Cars: Past, Present and Future - A Review of the Developments in the Last Century, the Present Scenario and the Expected Future of Autonomous Vehicle Technology. In *12th International Conference on Information in Control, Automation and Robotics* (hal. 191–198). https://doi.org/10.5220/0005540501910198

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bowser, A., Hansen, D., & Preece, J. (2013). Summary for Policymakers. In *Climate Change 2013 - The Physical Science Basis* (hal. 1–30). https://doi.org/10.1017/CBO9781107415324.004

Burrell, J. (2016). How the Machine “Thinks:” Understanding Opacity in Machine Learning Algorithms. *Big Data & Society*, (June), 1–12. https://doi.org/10.1177/2053951715622512

Buxton, R. T., Brown, E., Sharman, L., Gabriele, C. M., & McKenna, M. F. (2016). Using bioacoustics to examine shifts in songbird phenology. *Ecology and Evolution*, *6*(14), 4697–4710. https://doi.org/10.1002/ece3.2242

Cadwalladr, C. (2018). ‘I made Steve Bannon’s psychological warfare tool’: Meet the data war whistleblower. *The Guardian*. Taken from https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump

Chaloupka, M., Kamezaki, N., & Limpus, C. (2008). Is climate change affecting the population dynamics of the endangered Pacific loggerhead sea turtle? *Journal of Experimental Marine Biology and Ecology*, *356*(1–2), 136–143. https://doi.org/10.1016/j.jembe.2007.12.009

Chandler Lennon. (2019). Audubon Invites You to Join the 119th Annual Christmas Bird Count | Audubon. Diambil 9 April 2019, from https://www.audubon.org/news/audubon-invites-you-join-119th-annual-christmas-bird-count

Cigliano, J. A., Meyer, R., Ballard, H. L., Freitag, A., Phillips, T. B., & Wasser, A. (2015). Making marine and coastal citizen science matter. *Ocean and Coastal Management*, *115*, 77–87. https://doi.org/10.1016/j.ocecoaman.2015.06.012

Citizen Science Alliance. (2011). Citizen Science Alliance. Diambil 9 April 2019, from https://www.citizensciencealliance.org/philosophy.html%0Ahttp://www.citizensciencealliance.org/

Cocucci, A., Alcántara, J. M., Traveset, A., Arroyo, J., Navarro, L., García, D., Verdú, M. (2014). Beyond species loss: the extinction of ecological interactions in a changing world. *Functional Ecology*, *29*(3), 299–307. https://doi.org/10.1111/1365-2435.12356

Conrad, C. C., & Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: Issues and opportunities. *Environmental Monitoring and Assessment*, *176*(1–4), 273–291. https://doi.org/10.1007/s10661-010-1582-5

Courter, J. R., Johnson, R. J., Stuyck, C. M., Lang, B. A., & Kaiser, E. W. (2013). Weekend bias in Citizen Science data reporting: Implications for phenology studies. *International Journal of Biometeorology*, *57*(5), 715–720. https://doi.org/10.1007/s00484-012-0598-7

Crowston, K., Wiggins, A., Newman, G., Newman, S., Crall, A., & Graham, E. (2012). The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, *10*(6), 298–304. https://doi.org/10.1890/110294

Danielsen, F., Balete, D. S., Poulsen, M. K., Enghoff, M., Nozawa, C. M., & Jensen, A. E. (2000). *A simple system for monitoring biodiversity in protected areas of a developing country*. *Biodiversity and Conservation* (Vol. 9). Taken from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.499.6163&rep=rep1&type=pdf

Dearden, P., & Lunn, K. E. (2006). Monitoring small-scale marine fisheries: An example from Thailand’s Ko Chang archipelago Optimal Strategies for Marine Wildlife Tourism in Small Islands (PHD) View project Evaluating management effectiveness: Indicators for Marine Protected Areas in British Columbia View project Monitoring small-scale marine fisheries: An example from Thailand’s Ko Chang archipelago. *Fisheries Research*, *77*, 60–71. https://doi.org/10.1016/j.fishres.2005.08.009

Devictor, V., Whittaker, R. J., & Beltrame, C. (2010). Beyond scarcity: Citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, *16*(3), 354–362. https://doi.org/10.1111/j.1472-4642.2009.00615.x

Di Minin, E., Fink, C., Tenkanen, H., & Hiippala, T. (2018). Machine learning for tracking illegal wildlife trade on social media. *Nature Ecology & Evolution*, *2*(March), 406–407. https://doi.org/10.1038/s41559-018-0466-x

Dodge, S., & Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. *2017 26th International Conference on Computer Communications and Networks, ICCCN 2017*. https://doi.org/10.1109/ICCCN.2017.8038465

Edgar, G. J., Barrett, N. S., & Morton, A. J. (2004). Biases associated with the use of underwater visual census techniques to quantify the density and size-structure of fish populations. *Journal of Experimental Marine Biology and Ecology*, *308*(2), 269–290. https://doi.org/10.1016/j.jembe.2004.03.004

Engel, S. R., & Voshell, J. R. (2002). Volunteer biological monitoring: Can it accurately assess the ecological condition of streams? *American Entomologist*, *48*(3), 164–177. https://doi.org/10.1093/ae/48.3.164

Essl, F., Dullinger, S., Rabitsch, W., Hulme, P. E., Pyšek, P., Wilson, J. R. U., & Richardson, D. M. (2015). Delayed biodiversity change: No time to waste. *Trends in Ecology and Evolution*. https://doi.org/10.1016/j.tree.2015.05.002

Evans, C., Abrams, E., Reitsma, R., Roux, K., Salmonsen, L., & Marra, P. P. (2005). The Neighborhood Nestwatch Program: Participant Outcomes of a Citizen-Science Ecological Research Project. *Conservation Biology*, *19*(3), 589–594. https://doi.org/10.1111/j.1523-1739.2005.00s01.x

Farmer, R. G., Leonard, M. L., & Horn, A. G. (2012). Observer effects and avian-call-count survey quality: Rare-species biases and overconfidence. *The Auk*, *129*(1), 76–86. https://doi.org/10.1525/auk.2012.11129

Fattorini, S., Di Lorenzo, T., & Galassi, D. M. P. (2018). Earthquake impacts on microcrustacean communities inhabiting groundwater-fed springs alter species-abundance distribution patterns. *Scientific Reports*, *8*(1), 1501. https://doi.org/10.1038/s41598-018-20011-1

Fish, M. R., C ˆ Otéot, I. M., Gill, J. A., Jones, A. P., Renshoff, S., & Watkinson, A. R. (2005). *Predicting the Impact of Sea-Level Rise on Caribbean Sea Turtle Nesting Habitat*. *Conservation Biology* (Vol. 19). Taken from https://research.fit.edu/media/site-specific/researchfitedu/coast-climate-adaptation-library/latin-america-and-caribbean/eastern-carib-amp-dutch-antilles/Fish-et-al.--2004.--Bonaire-Sea-Turtle-Nesting--SLR..pdf

Galaz, V., Crona, B., Daw, T., Bodin, Ö., Nyström, M., & Olsson, P. (2010, Maret 1). Can web crawlers revolutionize ecological monitoring? *Frontiers in Ecology and the Environment*. John Wiley & Sons, Ltd. https://doi.org/10.1890/070204

Ganzevoort, W., van den Born, R. J. G., Halffman, W., & Turnhout, S. (2017). Sharing biodiversity data: citizen scientists' concerns and motivations. *Biodiversity and Conservation*, *26*(12), 2821–2837. https://doi.org/10.1007/s10531-017-1391-z

Geoghegan, H., Dyke, A., Pateman, R., West, S., & Everett, G. (2016). *Understanding Motivations for Citizen Science. Final Report on behalf of the UKEOF, University of Reading, Stockholm Environment Institute (University of York) and University of the West of England*. Taken from www.ukeof.org.uk

Hastie, T., & Tibshirani, R. (2007). Generalized Additive Models. *Statistical Science*, *1*(3), 314–318. https://doi.org/10.1214/ss/1177013609

Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (2018). *Citizen Science: Innovation in Open Science, Society and Policy*. *Citizen Science*. https://doi.org/10.14324/111.9781787352339

Hendry, A. P., Gotanda, K. M., & Svensson, E. I. (2017). Human influences on evolution, and the ecological and societal consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1712). https://doi.org/10.1098/rstb.2016.0028

International Telecommunication Union. (2017). *Measuring the Information Society Report 2017 Volume 1* (Vol. 1). Taken from https://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2017/MISR2017_Volume1.pdf

Irwin, A. (1995). *Citizen science: a study of people, expertise and sustainable development*. (Routledge, Ed.), *Choice Reviews Online* (Vol. 34). Environment and society. https://doi.org/10.5860/choice.34-0267

Jarvis, R. M., Breen, B. B., Krägeloh, C. U., & Billington, D. R. (2015). Citizen science and the power of public participation in marine spatial planning. *Marine Policy*, *57*, 21–26. https://doi.org/10.1016/j.marpol.2015.03.011

Johnson, M. F., Hannah, C., Acton, L., Popovici, R., Karanth, K. K., & Weinthal, E. (2014). Network environmentalism: Citizen scientists as agents for environmental advocacy. *Global Environmental Change*, *29*, 235–245. https://doi.org/10.1016/j.gloenvcha.2014.10.006

Keough, H. L., & Blahna, D. J. (2006). Achieving integrative, collaborative ecosystem management. *Conservation Biology*. https://doi.org/10.1111/j.1523-1739.2006.00445.x

Ko Lwin, K., & Murayama, Y. (2011). Web-Based GIS System for Real-Time Field Data Collection Using a Personal Mobile Phone. *Journal of Geographic Information System*, *3*, 382–389. https://doi.org/10.4236/jgis.2011.34037

Kobori, H., Dickinson, J. L., Washitani, I., Sakurai, R., Amano, T., Komatsu, N., Miller-Rushing, A. J. (2016). Citizen science: a new approach to advance ecology, education, and conservation. *Ecological Research*, *31*(1), 1–19. https://doi.org/10.1007/s11284-015-1314-y

Kotsiantis, S. B. (2007). *Supervised Machine Learning: A Review of Classification Techniques*. *Informatica* (Vol. 31). Taken from http://www.informatica.si/index.php/informatica/article/viewFile/148/140

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lepczyk, C. A. (2005). Integrating published data and citizen science to describe bird diversity across a landscape. *Journal of Applied Ecology*, *42*, 672–677. https://doi.org/10.1111/j.1365-2664.2005.01059.x

Lewison, R. L., & Crowder, L. B. (2007). Putting longline bycatch of sea turtles into perspective. *Conservation Biology*. https://doi.org/10.1111/j.1523-1739.2006.00592.x

Linda E. Kruger, M. A. S. (2002). Getting to Know Ourselves and Our Places Through Participation in Civic Social Assessment. *Society & Natural Resources*, *13*(5), 461–478. https://doi.org/10.1080/089419200403866

Lindenmayer, D. B., & Likens, G. E. (2009). Adaptive monitoring: a new paradigm for long-term research and monitoring. *Trends in Ecology and Evolution*, *24*(9), 482–486. https://doi.org/10.1016/j.tree.2009.03.005

Madin, J., de Moel, H., Vermaat, J. E., McClanahan, T., Maina, J., & Zinke, J. (2013). Human deforestation outweighs future climate change impacts of sedimentation on coral reefs. *Nature Communications*, *4*(1), 1986. https://doi.org/10.1038/ncomms2986

Mancini, A., & Koch, V. (2009). Sea turtle consumption and black market trade in Baja California Sur, Mexico. *Endangered Species Research*, *7*(1), 1–10. https://doi.org/10.3354/esr00165

Mason, M. (2010). Information disclosure and environmental rights: The Aarhus Convention. *Global Environmental Politics*, *10*(3), 10–31. https://doi.org/10.1162/GLEP_a_00012

Matzke, N., Ferrer, E. A., Barnosky, A. D., Wogan, G. O. U., Mersey, B., Marshall, C., Swartz, B. (2011). Has the Earth's sixth mass extinction already arrived? *Nature*, *471*(7336), 51–57. https://doi.org/10.1038/nature09678

Meeker, M. (2018). *Internet Trends Report 2018*. Taken from https://www.kleinerperkins.com/files/INTERNET_TRENDS_REPORT_2018.pdf

Nielson, J. (2006). Participation Inequality: The 90-9-1 Rule for Social Features. *Nielsen Norman Group*. Taken from https://www.nngroup.com/articles/participation-inequality/

Nov, O., Naaman, M., & Ye, C. (2008). What drives content tagging (hal. 1097–1100). https://doi.org/10.1145/1357054.1357225

Novak, J. (2009). *Mine, yours... ours? Designing for principal-agent collaboration in interactive value creation MINE, YOURS…OURS? DESIGNING FOR PRINCIPAL-AGENT COLLABORATION IN INTERACTIVE VALUE CREATION*. Taken from http://www.zora.uzh.chhttp//www.zora.uzh.chhttp://www.zora.uzh.ch

Obee, B., & Ellis, G. M. (1992). *Guardians of the whales : the quest to study whales in the wild*. (Alaska Northwest Books, Ed.). Whitecap Books. Taken from https://books.google.pt/books?id=XQ4XAQAAIAAJ&q=Guardians+of+the+Whales:+The+Quest+to+Study+Whales+in+the+Wild.+North+Vancouver,+British+Columbia:+Whitecap+Books&dq=Guardians+of+the+Whales:+The+Quest+to+Study+Whales+in+the+Wild.+North+Vancouver,+British+C

Oliver, T. H., Isaac, N. J. B., August, T. A., Woodcock, B. A., Roy, D. B., & Bullock, J. M. (2015). Declining resilience of ecosystem functions under biodiversity loss. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms10122

Papworth, S. K., Rist, J., Coad, L., & Milner-Gulland, E. J. (2009). Evidence for shifting baseline syndrome in conservation. *Conservation Letters*. https://doi.org/10.1111/j.1755-263X.2009.00049.x

Pauliuk, S., Majeau-Bettez, G., Mutel, C. L., Steubing, B., & Stadler, K. (2015). Lifting Industrial Ecology Modeling to a New Level of Quality and Transparency: A Call for More Transparent Publications and a Collaborative Open Source Software Framework. *Journal of Industrial Ecology*, *19*(6), 937–949. https://doi.org/10.1111/jiec.12316

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Perrot, E. (2011). Scikit-learn: Machine Learning in Python Fabian. *Journal ofMachine Learning Research*, *12*, 2825–2830.

Pierre, L., Dale, M. R. T., Marie-Josee, F., Jessica, G., Michael, H., & Donald, M. (2002). The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, *25*(March), 601–615.

Pimm, S. L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., … Loarie, S. (2015). Emerging Technologies to Conserve Biodiversity. *Trends in Ecology and Evolution*. https://doi.org/10.1016/j.tree.2015.08.008

Reisser, J., Proietti, M., Kinas, P., & Sazima, I. (2008). Photographic identification of sea turtles: Method description and validation, with an estimation of tag loss. *Endangered Species Research*, *5*(1), 73–82. https://doi.org/10.3354/esr00113

Ross, S. (1962). Scientist: The story of a word. *Annals of Science*, *18*(2), 65–85. https://doi.org/10.1080/00033796200202722

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Jacobs, D. (2012). Dynamic changes in motivation in collaborative citizen-science projects (hal. 217). https://doi.org/10.1145/2145204.2145238

Royle, J. A. (2013). *Review of: Mixed Effects Models and Extensions in Ecology with R*. Taken from http://arxiv.org/abs/1305.6995

Royle, J. A., Kery, M., Gautier, R., & Schmid, H. (2007). Hierarchical Spatial Models of Abundance and Occurrence From Imperfect Survey Data. *Ecological Monographs*, *77*(3), 465–481.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Schipper, J., Chanson, J. S., Chiozza, F., Cox, N. A., Hoffmann, M., Katariya, V., Young, B. E. (2008). The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science*, *322*(5899), 225–230. https://doi.org/10.1126/science.1165115

Schwarz, C. J. (2013). (R) Sampling, Regression, Experimental Design and Analysis for Environmental Scientists, Biologists, and Resource Managers. *The big R book*, 1668. Taken from http://people.stat.sfu.ca/~cschwarz/Stat-650/Notes/PDFbigbook-R/PDFbigbook-R.pdf

Science Communication Unit, U. of the W. of E. B. (2013). (2013). Science for Environment Policy In-Depth Report: Environmental Citizen Science. *Report Produced for the European Commission DG Environment*, (9), 32. https://doi.org/10.1007/978-94-007- 4587-2_7

Secord, J. A. (2005). Knowledge in Transit. *Isis*, *95*(4), 654–672. https://doi.org/10.1086/430657

Singh, J. S. (2002). The biodiversity crisis: a multifaceted review. *Current Science*, *82*(6). Taken from http://repository.ias.ac.in/72925/1/72925.pdf

Snäll, T., Kindvall, O., Nilsson, J., & Pärt, T. (2011). Evaluating citizen-based presence data for bird monitoring. *Biological Conservation*, *144*(2), 804–810. https://doi.org/10.1016/j.biocon.2010.11.010

Tang, L., & Shao, G. (2015). Drone remote sensing for forestry research and practices. *Journal of Forestry Research*. https://doi.org/10.1007/s11676-015-0088-y

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6). https://doi.org/10.1371/journal.pone.0021101

Thomaz, A. L., & Breazeal, C. (2006). Transparency and Socially Guided Machine Learning. In *International Conference on Development and Learning*.

Turnhout, E., & Boonman-Berson, S. (2011). Databases, scaling practices, and the globalization of biodiversity. *Ecology and Society*, *16*(1). https://doi.org/10.5751/ES-03981-160135

Turrini, T., Dörler, D., Richter, A., Heigl, F., & Bonn, A. (2018). The threefold potential of environmental citizen science - Generating knowledge, creating learning opportunities and enabling civic participation. *Biological Conservation*, *225*, 176–186. https://doi.org/10.1016/j.biocon.2018.03.024

Van Tienhoven, A. M., Den Hartog, J. E., Reijns, R. A., & Peddemors, V. M. (2007). A computer-aided program for pattern-matching of natural marks on the spotted raggedtooth shark Carcharias taurus. *Journal of Applied Ecology*, *44*(2), 273–280. https://doi.org/10.1111/j.1365-2664.2006.01273.x

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, *48*, 144–156. https://doi.org/10.1016/j.jmsy.2018.01.003

Whitelaw GS, Vaughan H, Craig B, A. D. (2003). Establishing Environmental, the Canadian Community Monitoring Network. *Environmental Monitoring and Assessment*, *Envi*(88), 409–418. Taken from http://cyber.sci-hub.tw/MTAuMTAyMy9hOjEwMjU1NDU4MTMwNTc=/10.1023%40a%3A1025545813057.pdf

Xu, T., Jin, X., Huang, P., Zhou, Y., Lu, S., Jin, L., & Pasupathy, S. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *TensorFlow: A System for Large-Scale Machine Learning* (hal. 619). Taken from https://tensorflow.org.

Young, B. E., Waller, R. W., Chanson, J. S., Stuart, S. N., Cox, N. A., Fischman, D. L., & Rodrigues, A. S. L. (2004). Status and Trends of Amphibian Declines and Extinctions Worldwide. *Science*, *306*(5702), 1783–1786. https://doi.org/10.1126/science.1103538