Master Graduation Report

Enabling Human-In-The-Loop Interpretability Methods of Machine Learning Models:

The Case of Bird Species Identification

TU Delft

Wei Zeng

November, 2021

TU Delft

Master Thesis
TU Delft, November 2021

Enabling Human-In-The-Loop Interpretability Methods of Machine Learning Models: The Case of Bird Species Identification

Supervisory team

Project chair
Alessandro Bozzon

Project mentor
Dave Rust-Murray

Collaboration support

Agathe Balayn

Nataša Rikalo

Author
Wei Zeng
MSc. Design for Interaction

Contact
zengweight@gmail.com

Acknowledgement.

Arriving at the end of this graduation project, I feel so lucky and honored to have the support from my supervisory team along the journey.

Agathe, thank you for bringing up this interesting research topic, and leading me along the way by generously sharing your domain knowledge and providing me with in-time technical support. The area of explainable AI was totally new to me before this project, but with your help and advice, I was finally able to design for it. And I think it was a enjoyable and exciting experience to be working with you.

Nataša, thank you for the brainstorms and ideas you provided me at the beginning of this project. And thank you for encouraging me to take this challenge. You have helped me a lot in founding this project.

Thank you, Dave, for always being so encouraging and supportive, and for mentoring me with your vast knowledge in both data science and design. I admire how you can always pinpoint the root of a problem and provide me with sound and feasible advice.

Thank you, Alessandro, for your serious remarks and attempts to push me to a greater level of understanding of the investigated topic. Because of you, I was able to get more out of this project than I had anticipated.

And I want to thank all the people that have participated in this project, including participants of the user tests and online surveys, as well as my dear friends who helped distributing the surveys or reaching out to the bird hobbyists for me. Thank you, Zhou Zichen, Li Guo, Mei Jian, Li Jingyi, Xu Mingyang, Liang Zhijian, Karen, Xiaoye, Pleun, Jiyoun, Li Chengtian, Bao Huien, Layla, Zheng Zijue, and those many netizens on Reddit and Douban who I don't know the names of, but have shared your insightful opinions. As a citizen science project, you helped me realize how powerful the combined power of general citizens can be. Thank you everyone so much; without your help, the project would not be possible.

Looking back on the previous two years of my master's trip, it was surely a challenging tour, with exotic cultures to embrace, outstanding people to work with, fresh new knowledge and research methodologies to absorb, and all kinds of daily challenges to overcome. I owe a huge thank you to everyone who came out over the last two years, from whom I have gained generous help and learnt a lot. Thank you for your company and for making my master's life a memorable experience.

Finally, I want to thank myself for the courage with which I faced problems and the perseverance with which I tried and learned from all of the defeats. Among all the things I acquired at IDE from the staff and my peers, the virtue of making mistakes and learning from them was the most essential. And I'll remember to carry it to the future path of my life.

Abstract.

To study how to involve the end-users in the development of machine learning explainability, this project has chosen the context of bird species identification. It intends to develop a platform where the end-users can learn bird knowledge while contributing to building the explainability of machine learning models. Among all the methods that equips machine learning models with explainability, this project adopts a framework called SECA (Semantic Concept Extraction and Analysis). In this framework, we require human-made-annotations to be made to the saliency maps of training photos to provide semantically understandable explanations to the end-users. On the other hand, we hope that the process of making annotations will also benefit the human annotators' skills in bird species identification, in order to motivate their participation.

Two main goals of the user research were: to understand the users' needs for learning and to know their capability in making the annotations needed by the project owners.

The user research started with qualitative and quantitative research to understand the current practices of the bird hobbyists, to define the target user groups, which were the birders with zero or little expertise.

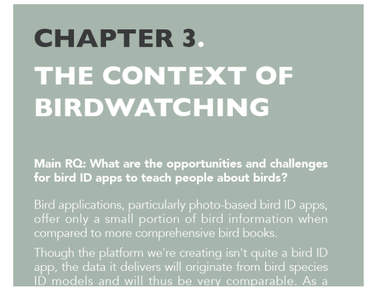
Then, in order to link their learning needs to the capability of machine learning explanations, three prototypes were built to collect their feedback. It was found out that they didn't care much about the justification or transparency of bird ID apps, compared to learning knowledge in distinguishing birds. Then came the annotation test when we found the participants were able to finish the annotation task with high correctness (>93% on average). And the most popular annotations of each task were 100% correct.

Finally, we built a functional, high-fidelity prototype with experiential interfaces and interactions, and tested it among 3 of the target users. They had positive feedback on the prototype's usability and the overall workflow, which proved the feasibility of our concepts. Recommendations on usability were drawn at the end of this test. Throughout the research and design phases in this project, we have developed an approach to involve end-users in the annotation process of an explainable bird species identification model for their own benefit of fun and learning, which could potentially be applied to broader deployments.

Reading Guide.

1. Chapter overview, summing ups and takeaways

Each chapter begins with a Chapter name, a main research question and an overview, and ends with a summing up and several takeaways.


CHAPTER 3.
THE CONTEXT OF BIRDWATCHING

Main RQ: What are the opportunities and challenges for bird ID apps to teach people about birds?

Bird applications, particularly photo-based bird ID apps, offer only a small portion of bird information when compared to more comprehensive bird books.

Though the platform we're creating isn't quite a bird ID app, the data it delivers will originate from bird species ID models and will thus be very comparable. As a

Summing up Chapter 1

This chapter addressed the opportunities and challenges that AI faced...

Takeaways

- AI technologies could be powerful tools in bird identification...
- We chose the SECA framework as a method to equip ML bird ID models ...

2. Research questions

The main research question of each Chapter will be further broken down into several sub-questions to investigate, which will be shown in this way.

A sub question may be followed by a supplementle question, indicating that the questions are similar by nature, differing only in focus, and can be answered in one sitting.

RQ10: To what extent are the end-users able to make annotations correctly on the photos with bounding boxes?

a. To what extent are the end-users able to make annotation correctly on photos that the model found hard to classify?

3. Highlights

Important knowledge or insights will be highlighted in this way.

4. Quotes

"The participants' statements will be quoted in this way." -Number of the participant who is quoted

Table of Contents.

Terminology.

Algorithms

A set of rules that a machine follows to achieve a particular goal.

Artificial intelligence(AI)

The theory and development of computer systems able to perform tasks that normally require human intelligence.

Birding, or bird-watching

Birding, or birdwatching, is the activity of observing wild birds in their natural habitats for scientific, conservation, recreational, and/or competitive purposes. (Moscovitch, 2019)

Birders, or bird watchers

The term used to describe the person who seriously pursues the hobby of birding. Maybe professional or amateur. (Birding, Volume 1, No.2)

Bird identification(Bird ID)

Identify the bird species with any tools or methods.

Crowdsourcing

Using the internet to attract and divide work between participants to achieve a cumulative result, however it may not always be an online activity

Computer Vision(CV)

A field of computer science that deals with gaining understanding and insights from digital images and videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do (Sonka, Hlavac, & Boyle, 2014)

Explainable Artificial Intelligence(XAI)

The ability of algorithms to explain their reasoning and characterize the strengths and weaknesses of their decision-making process

Interpretability/explainability

The degree to which a human can understand the cause of a decision.(Miller, 2019) The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made.

Machine learning(ML)

A set of methods that enable computers to learn from data and make predictions. (Molnar & Christoph, 2019)

Apps

Applications

RQ

Research questions

CHAPTER 1.
GENERAL INTRODUCTION

To start with, this chapter will give a general overview of this project and its context. Then, it defines the research questions and stakeholders of this project, after which, it anticipates the project planning and the final deliverables.

1.1 Birdwatching as a practice

Birding, or birdwatching, is the activity of observing wild birds in their natural habitats for scientific, conservation, recreational, and/or competitive purposes. (Moscovitch, 2019) The word birder is used to refer to people who seriously pursue the hobby of birding, no matter professional or amateur.

The term "bird watching" first appeared in a book called "Bird Watching" by Edmund Selous in 1901. The invention of optical equipment such as binoculars in the twentieth century allowed people to observe birds from a distance, paving the way for modern bird watching. (Moss, 2013)

Besides observing birds solely, some birders may expand their interest by taking classes, or joining local clubs to go for walks with other birders. (McIntosh, 2014)

Moreover, the birding practices could mean not only recreation for bird hobbyists but also an important part of ecological research. For example, the Christmas Bird Count (CBC) in the U.S collects bird reports from volunteered birders every year to monitor the population of birds for scientific or conservation purposes. (Wiersma, 2010)

1.2 Rise of computational tools for birdwatching and identification

Technology has changed the ways people learn, discuss and identify birds. Besides the traditional practices of checking bird books and joining local clubs, people now also seek to gain knowledge from online platforms and discuss with hobbyists in online communities. (McIntosh, 2014)

On professional websites like Birdwatching.com, Audubon.org, and the Cornell Lab of Ornithology website(www.birds.cornell.edu), people can find abundant educational information on birding, from how to distinguish birds to how to birding locations.

In online birding communities, people share their birding experience, observations, knowledge, and help each other with identifying unknown birds. Such communities include the iNaturalist website (iNaturalist.org), and some birding-related subreddits on Reddit. (www.reddit.com/r/birding).

When it comes to species identification, besides throwing the inquiries into one of the online communities, there are also platforms that guide people through the identification step by step. Such as the Bird watcher's digest (www.birdwatchersdigest.com), and the whatbird.com. In the WhatBird Wizard function on whatbird.com, by answering a series of questions like the spotted location, bill shape, wing shape, etc, the website will output a guess of the likely species.

1.3 Possibilities and challenges of working with AI models

Previous section we went through resources that are powered by the internet where birders can gain help with species identification.

Artificial intelligence will be another powerful tool to use, and will make the process more efficient.

If you post a photo of an unrecognized species on iNaturalist, it takes 18 days on average to get the correct answer from the community, with half of the inquiries being answered in the first 2 days. While with AI-enabled apps, it takes less than seconds. (iNaturalist Computer Vision Explorations · iNaturalist, n.d.)

Two main ways to identify bird species with AI are through the captured images, and through the recordings of their calls. In this project, we will only focus on image-based identification.

Some efforts have been made already by the data scientists towards applying AI in bird species identification.

One of the successful cases is the Merlin Bird ID app developed by the Cornell Lab of Ornithology (Figure 1-1). It can make identification with the pictures or audio recordings uploaded by birders. By February, 2021, it already had more than 611,000 active users. (Harrison, 2021)

Nevertheless, to be applied in species identification, AI still faces some challenges.

One of the main challenges it faces is the collection of numerous correctly-labeled image data, as it requires much expertise of the specific domain to know what category each bird image belongs to, and hiring domain experts is usually costly. (Van Horn et al., 2015)

The other challenge is the opaqueness of the AI's reasoning process, even to its creator, because of its complexity.

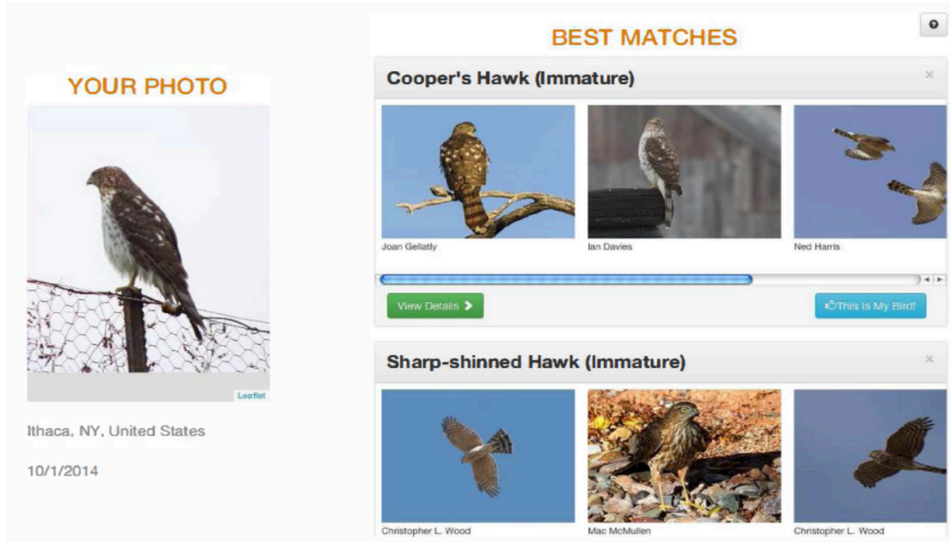


Figure 1-1. Screenshot of the Merlin Photo ID, a publicly available tool for bird species classification (Van Horn et al., 2015)

The word “black box” is used in machine learning to describe models that can not be understood by looking at their parameters, such as the neural network, which consists of a large number of interacting and nonlinear parts. (Molnar & Christoph, 2019)

State-of-art bird species identification tools based on image classification tasks mostly employ neural models. (Wäldchen & Mäder, 2018) This means that the model’s reasoning is hard for humans to understand, or debug, with the end-users not knowing when the prediction is trust-worthy and the developers not knowing whether and where it goes wrong.

1.4 Explainable AI and SECA as a potential future

To make the “black box” process understandable to the end-users, data scientists have been seeking ways to enable the model to explain its reasoning behind the decision. The research field explainable AI or XAI aims at increasing the transparency of AI systems and study the influence of the increased transparency on end-users (figure 1-2). (Anik & Bunt, 2021)

Examples of providing explanations are telling patients what are the indicative complaints of a certain diagnosis (Lundberg et al., 2018), or helping the factory workers to know the efficiencies in the production process (Dhurandhar, 2018).

Different types of problems will need different strategies for explaining. For the image classification tasks based on

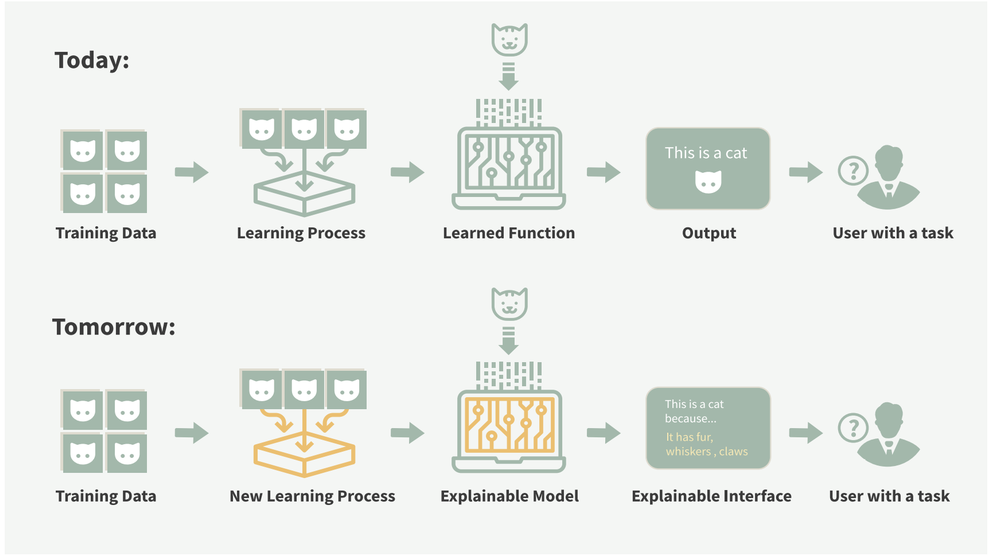


Figure 1-2. Vision of explainable AI

neural networks like in our case, saliency maps (i.e. a highlight of the most important pixels for the classification of a given image) can be used to reveal how each pixel contributes to a particular prediction. (Molnar & Christoph, 2019)

To make the saliency map understandable to end-users, a framework called SECA (Semantic Concept Extraction and Analysis) tried to produce semantically understandable explanations out of the highlighted pixels.

To provide such explanations, it applies a human-in-the loop method, which requires annotations made by humans to describe the concept of the highlighted areas in the saliency maps. (Figure 1-3)

Figure 1-3. Example of the saliency map(right) of a bird photo(original photo on the left), brighter pixels are more important for the classification

In order to apply this framework, we also need good reasons to motivate people’s participation in the annotation process. We assume that the annotation tasks could benefit the annotators themselves in their identification skills, which we will verify throughout the research of this project.

1.5 Stakeholders involved

1.5.1 Collaboration with SECA developers

This project adopts the idea of SECA (will be further elaborated in Chapter 2) and collaborates with the SECA developers. The developers support the execution of this project by providing real data materials generated by a bird species classification model, as well as giving technical advice on the design space around the ideas.

They will be considered as both the stakeholders and the owners of the conceived product.

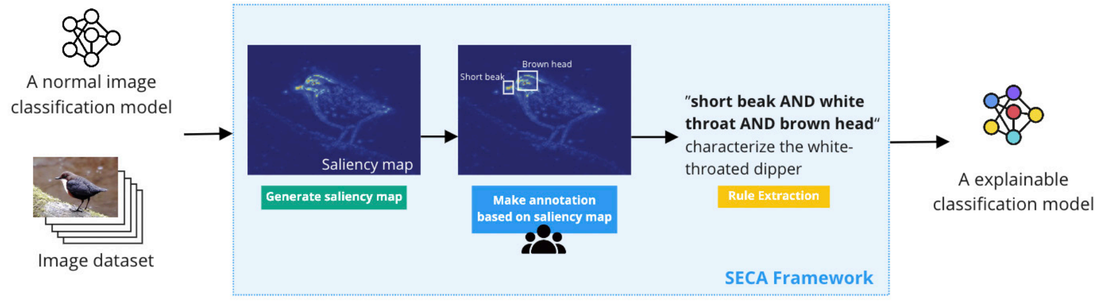


Figure 1-4. The workflow of SECA

1.5.2 The end-users

The other side of this project lies with the end-users, who could either be the birding hobbyists (birders) who take bird watching as a serious hobby, or generally someone who is interested in learning to tell birds apart. In the conceived app, they learn how to identify birds while contributing their annotations to the development of explainability.

Their expertise in telling birds apart ranges from the entrance to expert-level. In Chapter 4, it will be further defined which group of people will be the target users.

1.6 Problem definition

1.6.1 Project aim

The aim of this project is two-folded. One is to collect annotations from the end-users to improve the interpretability of a bird identification model. The other is to teach the end-users knowledge in identifying birds.

Project aim

1. Explore how the end-users can contribute to the development of machine learning interpretability.

2. Develop a bird species identification platform with explanations that engage the birders to use and enable their learning.

The first aim lies along with the interest of the project owners themselves, also the machine learning technologists'.

The second one lies with the interest of the bird hobbyists. While it's not in their interests to know the mechanism of the machine learning algorithms or to improve them, it is crucial for this project to find out how the explanation methods could satisfy their needs in order to incentivize their use.

We assume that the learning can happen in two ways, one is through looking at the explanations produced by the explainable model, one is through the process of making annotation itself (figure 1-5).

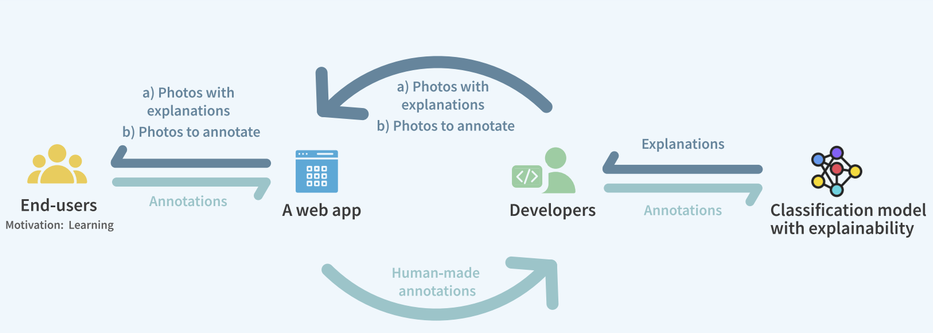


Figure 1-5. The information exchange flow between developers and the end-users

1.6.2 Main research questions

Research questions

1. Can the end-users learn about bird species identification with ML explanations and the annotation tasks?

2. How can we engage the end-users in contributing their annotations with the design of a game-like bird identification learning product?

3. Are the end-users able to make annotations needed by the developers through a game-like process?

It is worth mentioning that these research questions did not emerge from the start, but rather after several rounds of exploration and alterations. The reasons for the adjustment in research questions will be discussed in Chapter 5 (The Explanation Prototypes), where explanation prototypes were evaluated.

The main research questions were divided into several sub-questions to be explored in different chapters.

1.7 Project Planning

The project goes through the following 4 phases following the double diamond design approach (figure 1-6) ("The Double Diamond Design Process Model," 2005).

Discovery

During the discovery phase, qualitative research such as interviews will be carried out to find out what are the current practices of bird hobbyists to learn identifying birds. **(Chapter 3)**

> **RQ:** What are the opportunities and challenges of bird ID apps in teaching people to learn about birds?

Define

Online surveys will be done to learn about the learning habits and demands of different levels of bird lovers in order to define who we are designing for. The findings will determine the set of people we will design for, as well as their present practices. **(Chapter 4)**

> **RQ:** What are opportunities and challenges for different levels of birders to adopt bird ID apps as their learning tools?

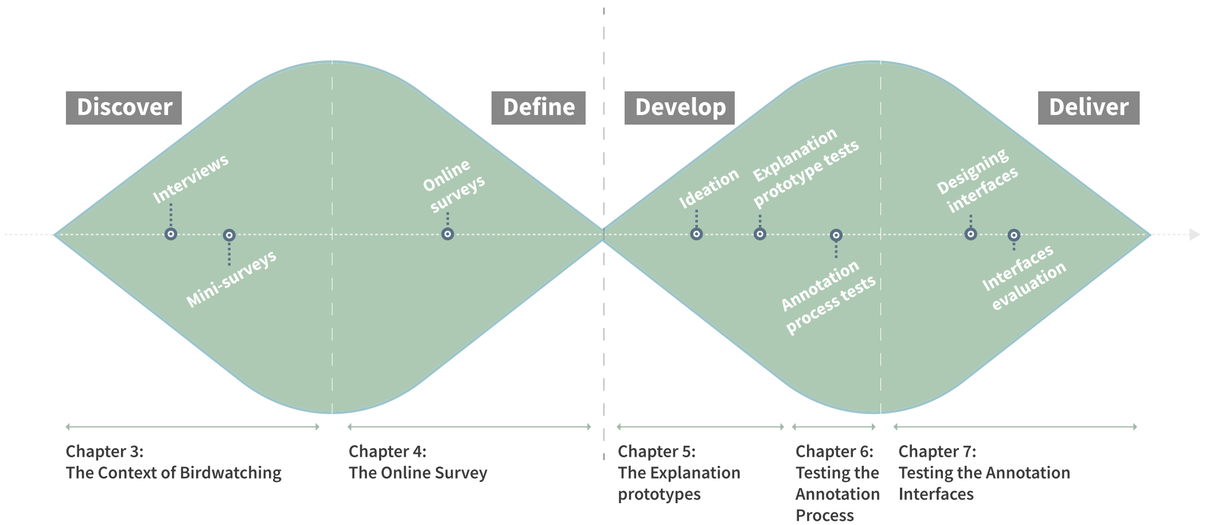


Figure 1-6. Overview of the project

Development

During the development phase, by testing out experiential prototypes among the target users, it will be explored how to link the technical capabilities for explanations and the developers' needs for annotation to the end user's needs.

There will be two rounds of user tests in total, one focusing on the explanation the other focusing on the annotation **(Chapter 5, Chapter 6)**

> **RQ:** What are the design opportunities for explainable AI models in the birding community?
>
> To what extent are the end-users able to make annotations correctly on the photos pre-processed by the developers?

Delivery

In this phase, hi-fi interfaces of the conceived product will be developed and tested among the target users to validate the feasibility of chosen concept, as well as to gain usability recommendations for future work. **(Chapter 7)**

> **RQ:** How to design an interface that people would like to use for learning and making annotations?

Finally, as conclusions of the entire project, discussions of the completed research activities, as well as the future possibilities we identified from the findings, will be drawn. **(Chapter 8)**

1.7 Outcomes

This is a research through design project, which means the design activities will finally lead to discussions in the chosen research field.

Firstly, based on user research of early stages, prototypes for a digital bird identification learning platform will be developed based on an interpretable bird identification model. Then, the prototypes will be utilized for several rounds of explorations and assessments throughout the project. Finally, at the conclusion of this project, we looked back at the conducted design research assignments, concluded the design implications, and answered all the main research questions posed.

In summary, the project made the following contributions:

- Exploring the opportunities of applying explainable AI in the education of bird ID knowledge
- Evaluating the idea of a game-like bird ID learning platform, which can collect annotations from the end-users, based on the SECA framework
- The findings demonstrated the feasibility of including general citizens in the development loop of an SECA-based explainability technique, as well as recommendations for future research.

Summing up Chapter 1

This chapter addressed the opportunities and challenges that AI faced when applied to the field of birdwatching, as well as our idea for how Explainable AI might be a solution for those challenges. Accordingly, research questions and design assignments were raised. In the next chapter, we'll go over the academic and technological background in greater depth to enable a better understanding of what the research issue entails.

Takeaways

- AI technologies could be powerful tools in bird identification, yet they can be perplexing due to their lack of transparency. Explainable AI (XAI) is a possible solution to its opaqueness.
- We chose the SECA framework as a method to equip ML bird ID models with explainability, where annotations are required from people to make the explanations semantically meaningful.
- We assume that learning would be the motivation for end-users to participate in the annotation process. And the learning could happen in seeing the produced explanation and in making annotations itself.

CHAPTER 2.
THE ACADEMIC BACKGROUND

To clarify the design scope and chosen research questions, this chapter presents an overview of the state-of-the-art academic background as well as some basic notions around the chosen context, such as explainable AI and citizen science.

2.1 Machine learning

Machine learning(ML) is a set of methods that enable computers to learn from data and make predictions. (Molnar & Christoph, 2019)

It is different from traditional programming methods, where specific rules were given by humans for the computers to produce output out of the input data. With machine learning, the computer learns the function itself from pairs of input and output examples. (Hohman, 2020)

For example, a machine learning model can learn from historical data of the house price in relation to influencing factors like size, location, floorplan and so on, and estimate a house price correlating to the newly input data.

Machine learning can be applied in various cases to make predictions, from estimating the house price, predicting weather to detecting street signs. The models development for these different scenarios follow a similar process (figure 2-1):

1) Data collection: get a dataset with the outcomes that we want to predict, and the paired information used to predict the outcome;

2) Training the model: entering the data into a ML algorithm to train the model;

3) Using the model: use the model with new data. (Molnar & Christoph, 2019)

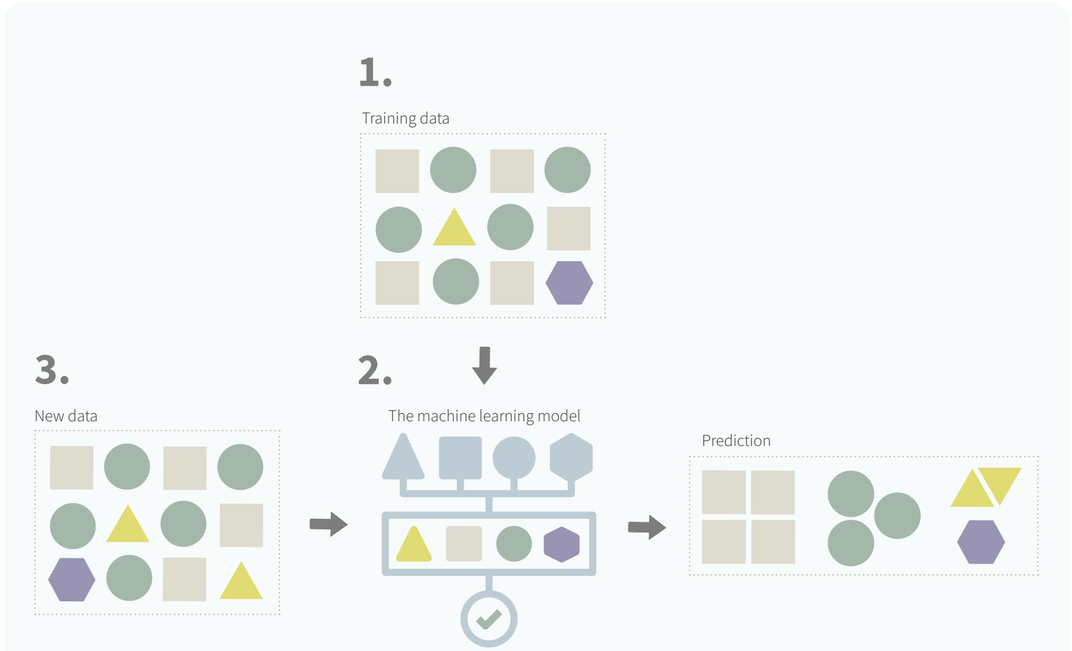


Figure 2-1. General process of utilizing machine learning

2.2 Computer vision

To enable image-based species classification, we need specific machine learning approaches called Computer Vision.

Computer Vision is a field that seeks to enable computers with the ability that human's visual systems can do (Sonka et al., 2014). It can deal with tasks like: image classification, object detection, face recognition, and so on.

A computer vision pipeline is made up of two phases: **feature extraction** and **classification**. (Wäldchen & Mäder, 2018)

A feature in machine learning is a measurable property or characteristic of a phenomenon being observed (Elgendy, 2020). In the **feature extraction** phase, raw input data will be transformed into meaningful features for a given classification task. For example, millions of pixels that make up an image, each with a color value, will be transformed into information like shape, texture, or color, to be useful for the classification problem.

Then, in the **classification** phase, each feature that comes out from the feature extraction will be mapped into a score of confidence through a classifier. And the score will be used for generating the prediction. (Wäldchen & Mäder, 2018)

In typical machine learning problems, the features are manually extracted by domain experts and then input into a classifier to predict the output, however in our case, where deep learning methods are employed, the feature extraction and classification tasks are done automatically end-to-end.

To be exact, the neural network automatically identifies features within the input image, and learns the importance of features through trials by attaching random weights to them and seeing how the output predictions are impacted. Then by comparing the scores of different species in the model, the model decides which species the bird in the input image is more likely to be.

Figure 2-2 shows the comparison between bird identification with human brains and that with computer vision.

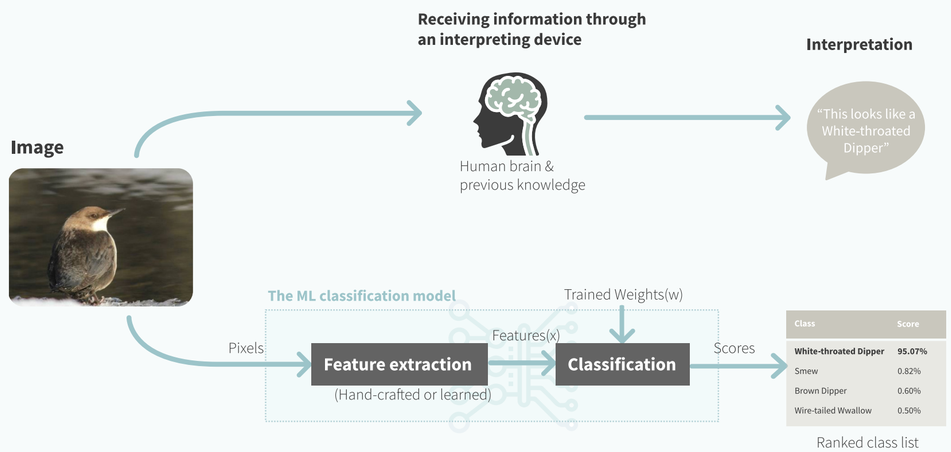


Figure 2-2. Bird species identification with human brains and computer vision

2.3 The Explainability of Machine Learning Models

2.3.1 Definition of concepts

While AI systems enabled by machine learning algorithms are often not transparent, the studies on explainable AI(XAI) seek ways to increase system transparency, and understand the impact of increased transparency on users' perceptions. (Anik & Bunt, 2021)

AI systems can be made understandable by providing machine learning explanations.

Miller (2017) has defined **explanation** this way: "An explanation is an answer to a why-question."(Miller et al., 2017)

In machine learning, **explanations** mean providing to the end-users textual or visual artifacts, explaining the relationship between the input features and the prediction result. (Ribeiro et al., 2016)

Figure 2-3 shows an example of providing explanations with visuals, where Ribeiro et al.(2016) tried to explain the top 3 predicted classes produced by an image classification model (Google's pre-trained Inception neural network). The top 3 predicted classes were "Electric guitar"(p=0.32), "Acoustic guitar"(0.24) and "Labrador"(0.21). The explanation was made by highlighting the pixels in the image that support each prediction (figure 2-3 (b)(c)(d)), so that the audience know the reasoning behind.

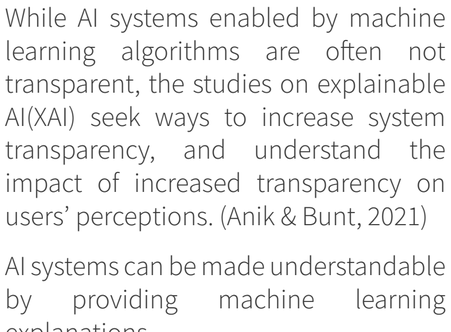


Figure 2-3. Explaining an image classification prediction. (Ribeiro, 2016)

2.3.2 Goals of machine learning explanations

Nothdurft (2013) concluded that there are overall five goals that a machine learning explanation can reach, which are: justification, transparency, relevance, conceptualization, and learning. These five goals are overlapped and interplay sometimes.

Goal of Explanation	Description
Justification	Explain the motives of the answer?
Transparency	How was the systems answer reached?
Relevance	Why is the answer a relevant answer?
Conceptualization	Clarify the meaning of concepts
Learning	Learn something about the domain

Figure 2-4. The different goals an explanation can pursue (Nothdurft et al., 2013)

This project was primarily structured around the goal of learning, but we weren't sure if it was the goal that our target customers were interested in. We'll also look into whether or not additional objectives are important. These objectives will be used as metrics in Chapter 5 to assess how well the explanation prototypes assist end-users in various ways.

2.4 Different types of explanations

There are different types of explanations for machine learning models, and they can be categorized with different criteria. For design ideation, it's important to learn what are the possible forms of explanations. As we want to study the application and impact of the SECA framework, we hope the ideas also fall into the scope of SECA's capability.

2.4.1 Taxonomies on the technical dimension

Technically speaking, an explainability method, according to how it was developed and functioned can be divided into intrinsic/post-hoc, local/global.

The **intrinsic/post-hoc dimension** refers to whether the explainability is accomplished by limiting the complexity of the machine learning model (intrinsic) or by using post-training analysis tools (post hoc). The **local/global dimension** distinguishes whether the output of this method explains a single prediction (local) or the overall model behavior (global). (Molnar & Christoph, 2019)

SECA is a post-hoc explainability method that could be applied to any existing classification models, as defined by the criteria above. Meanwhile, although SECA is a global explainability method, the global explanations it generates are drawn from local explanations.

We simply focus on the development of the local part of this project, i.e. explaining individual predictions, and we don't pay much attention to how it was transformed into global explainability. The mechanism behind will be further explained in section 2.5.

2.4.2 Taxonomies on the representation dimension

Apart from the technical dimension, I found the perspective of **explanation target** and **explanation medium** most relevant to designing the interaction and interfaces for explanations. The explanation target aspect determines the content of the explanation, while the explanatory medium determines the representation form of it.

From the perspective of the **explanation target**, according to Sokol & Flach (2020), explanations can be made on each component of the machine learning process, which is data, models and predictions. To explain data, in our case of image classification, we can show images in our dataset that are important or abnormal. To explain the model, we could extract and output the general rules of its function. While single predictions can be explained by revealing how particular data points have led to the prediction.

From the **explanation medium**'s perspective, an explanation can be in the form of statistics summarization, visualization, textualization, formal

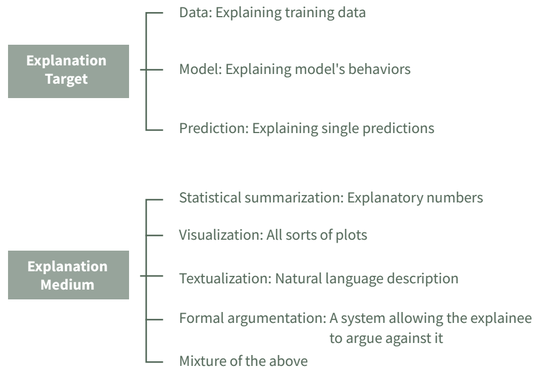


Figure 2-5. The explanation variants under the property of Explanation Target and Explanation Medium

argumentation, or a mixture of them. Examples of the first one could be coefficients of a linear model, or a statistical summary of the dataset. Visualization could be highlighting pixels on images. The textualization could be any explanation in the form of a textual description. (Sokol & Flach, 2020)

2.4.3 Applying to ideation

The nature of SECA will limit how the explanation could be developed and presented under this framework.

For example, the mechanism of SECA determines that neither the statistics summarization nor formal argumentation would be a proper way to deliver its outcome.

These variants of explanation systems will be the basis of the ideation in Chapter 5, and be evaluated among the target users.

2.5 Unpacking the SECA framework

In Chapter 1, we have briefly talked about the interpretability method called SECA (Semantic Concept Extraction and Analysis), which produces semantically understandable explanations for end-users, enabled by a human-in-the-loop method (Balayn et al., 2021). The following sections unpack the procedures of the SECA framework. During the project's duration, these SECA development steps will be tailored to the bird ID context and evaluated through the tests of interactive prototypes.

2.5.1 The development steps

In general, the SECA framework can be divided into two parts:

1.The local part, which collects annotations to describe highlighted areas on the training data;

2.The global part, in which local explanations are gathered and translated into final global explanations that explain the entire prediction class.

In detail, the complete process as follows will take place to get a model interpretability (figure 2-6):

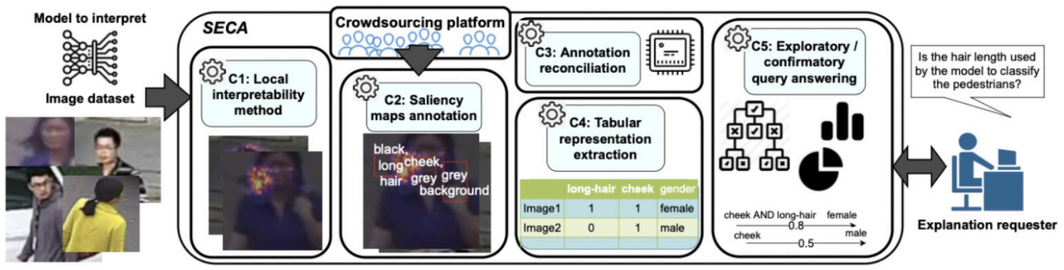


Figure 2-6. Overview of the SECA framework (Balayn et al., 2021)

Step 1 (C1): With the classification model to explain, generate saliency maps or specific images in the dataset.

Saliency maps highlight the pixels in the image that the model has found important for the classification. Using saliency maps helps generate relevant explanations and reduces the annotation effort.

Step 2 (C2): Have annotators drawing bounding boxes and labeling images.

The extraction of semantic concepts can not be automated currently, so human labor is needed for drawing bounding boxes and making annotations.

In the original setting of SECA, workers were asked to identify meaningful representations (objects) in the salient pixel area, and draw bounding boxes around that area. Then they were asked to describe the identified objects with a type word (eg. hair) and an attribute word (eg. long, black).

Step 3 (C3, C4, C5): Then, local explanations coming from the previous steps will be developed into global ones. Specifically, based on the annotation on single instances, the developers extract general rules for the classification, and make it into something accessible to the explainees, for example, a dialogue system that people can query.

With such explainability built, the explanations on classification models can be delivered to the end-users in different forms (see section 2.4.2), not limited to the dialogue system mentioned here.

As a result, the end-users will be provided with not only the prediction outcome (what species the target bird is) during their usage but also the explanation of that specific prediction(what features characterize that species).

2.5.2 The role of SECA in this project

Currently, the SECA method hasn't been evaluated with real end-users, and the output of SECA is not designed to allow usage by people without a computer science background. Besides, the annotations previously were collected from crowdsourcing platforms like Amazon Mechanical Turk (MTurk), instead of citizen scientists who have more or less interest and knowledge in the studied topic.

The developers of the SECA framework are now interested in knowing how to make the laymen in ML understand and benefit from the output of the ML interpretability, and also in how to effectively collect human-made annotations to improve the models.

This project will employ the local part of the SECA's workflow as the product's foundation. To collect annotations and produce local explanations, the platform's development will follow SECA's development flow (C1 and C2 in figure 2-6). Because it is more of a data science concern, the process of transforming local explainability into a global one would be excluded from the project's scope. However, we would investigate how the final global explanations could be conveyed to end users.

To summarize, the research area is how to provide an explanation with the given training data and how to build an annotation flow that is engaging and user-friendly for end-users.

2.6 The citizen science and citizen scientists

2.6.1 Definition of concepts

As introduced in Chapter 1, human annotators are needed in a SECA method to build an explainable identification model. In this project, we want to use **citizen scientists** to make the annotations needed.

Citizen scientists, by definition, are non-professional scientists or enthusiasts in a certain domain who voluntarily gather and/or process data as part of a scientific investigation in fields such as archeology, astronomy, and ecology, and others. (Silvertown, 2009; Van Horn et al., 2015)

A citizen science project is a project that involves citizen scientists. Nowadays, there are many projects that have been specifically designed or adapted for amateurs, to work on professional scientific topics.

2.6.2 Cases of citizen science projects

The birding community has a long history of contributing to citizen science projects.

For example, the Christmas Bird Count, being one of the longest-running wildlife census projects and still running today, was initiated by the National Audubon Society in December 1900. Every year during the Christmas period, the project collected bird-counting reports from volunteers for three weeks, to monitor the population of birds for scientific or conservation purposes. (McIntosh, 2014)

Similarly, eBird, a platform that allows people to document what bird species they have spotted at which location, has become a highly useful database for scientists. The collected observations from birders have been used for research such as the impact the extreme weather has on birds' distribution(Cohen et al., 2020), shifts in species' calls (Otter et al., 2020), changes in birds' seasonal distribution, etc (Sullivan et al., 2009).

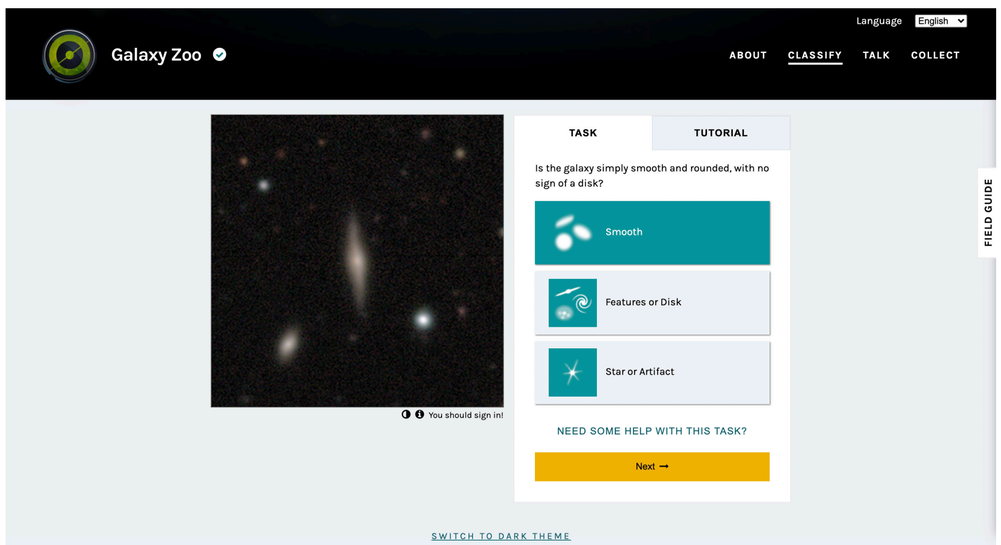


Figure 2-7. Screenshot of the Galaxy Zoo project webpage

Outside the birding field, Zooniverse (www.zooniverse.org/) is a citizen science web portal that hosts dozens of citizen science projects. The website grew from the original Galaxy Zoo project, lasting from 2007 to 2009, which invited people to help in the morphological classification of large numbers of galaxies. During the project's run, over 100,000 participants completed over 40 million classifications, averaging 38 classifications per galaxy.

In each of the projects on Zooniverse, users can view research data as photographs, video, and audio on one of the Zooniverse websites. And they were presented with a short guide or lesson on how to complete the required analysis to recognize, classify, mark, and label the data as researchers would. (Simpson et al., 2014)

By March 2019, Zooniverse already has 1.6 million registered volunteers. Astronomy, ecology, cell biology, humanities, and climate science are among the fields represented in the projects. (*Combining Artificial Intelligence and Citizen Science to Improve Wildlife Surveys*, 2019)

2.6.3 Strengths and drawbacks of employing citizen scientists

Compared to paid human labor recruiting from crowdsourcing platforms, there are two main reasons for recruiting citizen scientists for the annotation tasks in this project: lower cost and higher quality.

Researchers discovered that citizen scientists recruited at no cost are substantially more accurate than workers on MTurk (a crowdsourcing marketplace) in a study conducted by Cornell Lab of Ornithology, in which they produced a large-scale dataset for bird species identification with the help of citizen scientists. (Van Horn et al., 2015) This could be due to citizen scientists' greater skill and passion in this field, as well as the lack of spammers.

Nevertheless, the use of citizen scientists had several drawbacks, including a lesser volume of data collected and a longer period for researchers to identify ways to collaborate with different communities in this domain.

2.6.4 Motivations of the participants

The citizen scientists participate in those projects mostly for the benefit of getting knowledge from the volunteering experience, or for the benefit of the research project, ideally for both. (Silvertown, 2009)

For this project, we assume there are two ways that the bird ID learning platform could attract birders in using it, one is helping people to learn bird knowledge, the other is facilitating people's trust towards the prediction, which will be verified in the research (figure 2-8).

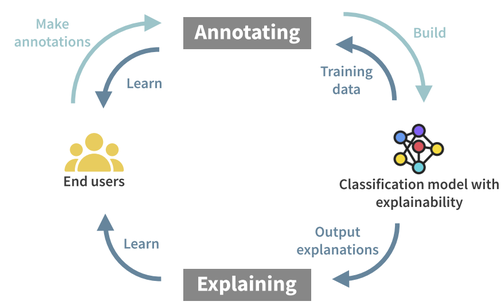


Figure 2-8. End-users learning through annotating and explaining process

Summing up Chapter 2

The state-of-the-art development process for machine learning and computer vision was introduced in this chapter. After that, it gives an overview of ML explainability examples and taxonomies. The SECA framework, which we used in this project as one of the ML explainability methodologies, was thoroughly explained. Finally, it discussed citizen science and the possibilities of launching a citizen science project in our environment.

Takeaways

- The image classification tasks are made possible with Computer Vision (CV), which generally consists of two steps: feature extraction and classification.
- Technologically speaking, SECA is a post-hoc, global explainability method. The global explanations it generates, on the other hand, are derived from the local ones, which are based on annotations of the highlighted areas. And this project focuses on the collection of annotations, which is the local part of the SECA framework.
- Though the overall development process has been framed, the delivery form of SECA's output varies, which offers flexibility for experimenting with how to communicate the generated explanations to end-users.
- Citizen science projects enlist volunteers to assist with scientific research, which is usually done at a low cost and with greater accuracy compared to human labor on crowdsourcing platforms. This project is envisioned as a citizen science project.

CHAPTER 3.

THE CONTEXT OF BIRDWATCHING

Main RQ: What are the opportunities and challenges for bird ID apps to teach people about birds?

Bird applications, particularly photo-based bird ID apps, offer only a small portion of bird information when compared to more comprehensive bird books.

Though the platform we're creating isn't quite a bird ID app, the data it delivers will originate from bird species ID models and will thus be very comparable. As a result, we'd like to know at this stage what knowledge is needed for birders to tell birds apart and what role the bird ID apps play in birders' education. Knowing this will give us insight into the features that can be expected while utilizing explainable bird ID applications as learning tools.

In this chapter, I delved into the world of bird-watching with qualitative research that included online interviews and mini-surveys with birders to learn about their present bird-watching routines.

A qualitative study was conducted to learn about people's experiences in the field of birding. The research is divided into two parts: semi-structured interviews and sending out queries and gathering replies in online birdwatching interest groups.

3.1 Background

The history of birding can be dated back to the 20th century.

Bird books, which date back to Gilbert White's Natural History of Selborne (1788) and John James Audubon's illustrated Birds of America (1827–38), inspired birding as a recreational activity and culminated in field guides like H.F. Witherby's five-volume Handbook of British Birds (1938–41) and Roger Tory Peterson's Field Guide to the Birds of the World (1947).

Like in many other domains, books are considered the most reliable resource to learn about birds. Bird Books with pictures and descriptions of bird species, often provide detailed information not only on the appearance of birds but also on their behaviors, distribution, tricks for identification, and so on.

Bird apps are also popular among birders nowadays, including but not limited to apps that serve as digital field guides (Audubon), identification tools (Merlin bird ID), birding record tools (eBird). Some apps belong to more than one type listed above. For example, most identification tools also provide an exhaustive list of local birds, as well as details such as photographs, behaviors, and habitats information, therefore can also be used as digital field guides.

Bird ID apps help people identify birds with different techniques, the most common ones of which are ID through photos, sound recordings, and through the users' descriptions of the bird's features (usually by asking users a few questions on size, shape, habitat, etc).

As the platform we are designing would use information similar to bird ID apps, we are curious how this information can help people in their learning route.

This project studies the explainability of image classification, so at this stage, we would like to know specifically about how the image-based bird ID apps are used by birders, what information they pay attention to, and what are the challenges of using them.

Research Questions

RQ1: What are the tools the birders currently use to learn about birds?

a. What knowledge do they learn with these tools?

RQ2: What do they find most helpful/challenging in their process of learning to identify birds?

RQ3: What do the birders currently use bird apps for?

3.2 Semi-structured interviews

To get some general knowledge of the context, I reached out to three birders to know their experience in learning around birds, and what they have found challenging or helpful throughout the ways.

3.2.1 Set-up

Research questions:

RQ1: What are the tools the birders used and currently use to learn about birds?

RQ2: What do they find most helpful/challenging in their learning process?

The main questions asked were:

1. How do you get started in identifying birds?

2. What helps you at different stages of learning?

3. Throughout the learning experience, which part did you find most helpful/challenging?

4. Have you used any bird ID app? What do you use it for?

And more detailed inquiries were brought up during the interviews around the main questions above.

3.2.2 Participants

3 birders participated in the interviews, with their experience in birding (counting from their first birding) and background shown in figure 3-1.

Participant NO.	Birding experience	Background
P1	2.5 years	Design
P2	5 years	Life science
P3	10 years	Life science

Figure 3-1. Overview of the interviewees

A summary of the semi-structured interview results has been documented in Appendix A.

3.3 Mini surveys

While the semi-structured interviews have provided insights on the birders' general birding experience, the aim of the pre-survey is to get in-depth insights on how the birders currently use bird apps, in a shorter time and of a larger range.

3.3.1 Set-up

Research questions:

RQ3: What do the birders currently use bird apps for?

The specific questions asked were:

1. Do you often use bird recognition apps?

2. What is the main purpose of using the bird recognition app (find the general direction/verify your guess/look for more information...)?

3. What information will you pay attention to when using the App?

4. Under what circumstances will the prediction result be considered credible/incredible for you?

Questions were sent out in the r/birding subreddit on Reddit, and birding groups on Douban (the Chinese version of Reddit).

3.3.2 Participants

10 participants in total responded to the queries, all of whom are people interested in bird watching with more or less birding experience.

Participant NO.	Channel
R1~R3	/whatsthisbird on Reddit
R4~R10	Birds group on Douban

Figure 3-2. Overview of replies for mini-surveys

The replies were labeled as R1~R10, with R1~R3 were in English, R4~R10 in Mandarin, which was translated into English for analysis.

All the original responses were documented in Appendix B.

3.3.3 Statement card analysis

Statement card analysis is a method for interpreting data and finding patterns out of the qualitative raw data by doing a clustering exercise. (Sleeswijk Visser et al., 2007)

In this analysis step, responses that were considered insightful were extracted and analyzed with the approach of statement card analysis (figure 3-3, Appendix C).

Specifically, the researcher picked out the statements from the raw data and concluded each of them into one or two sentences. Then, the statements were made into statement cards. Later, the researcher went through all the statement cards and made them into different clusters according to the themes they were talking about.

Cluster1: Recognizing errors in the app's identification

In this cluster, quotes were mainly about what information the participants would pay attention to and how they verified the correctness of the prediction results.

Cluster2: Overall attitude towards bird ID apps

In this cluster, statements were around people's attitude towards the bird ID apps in general, such as how they trust the predictions, and how they see the apps as part of their learning.

Cluster3: Bird ID apps in relation to other practices

This cluster reveals how the bird ID apps fit into their other birding practices, as well as what role they play in relation to other learning tools.

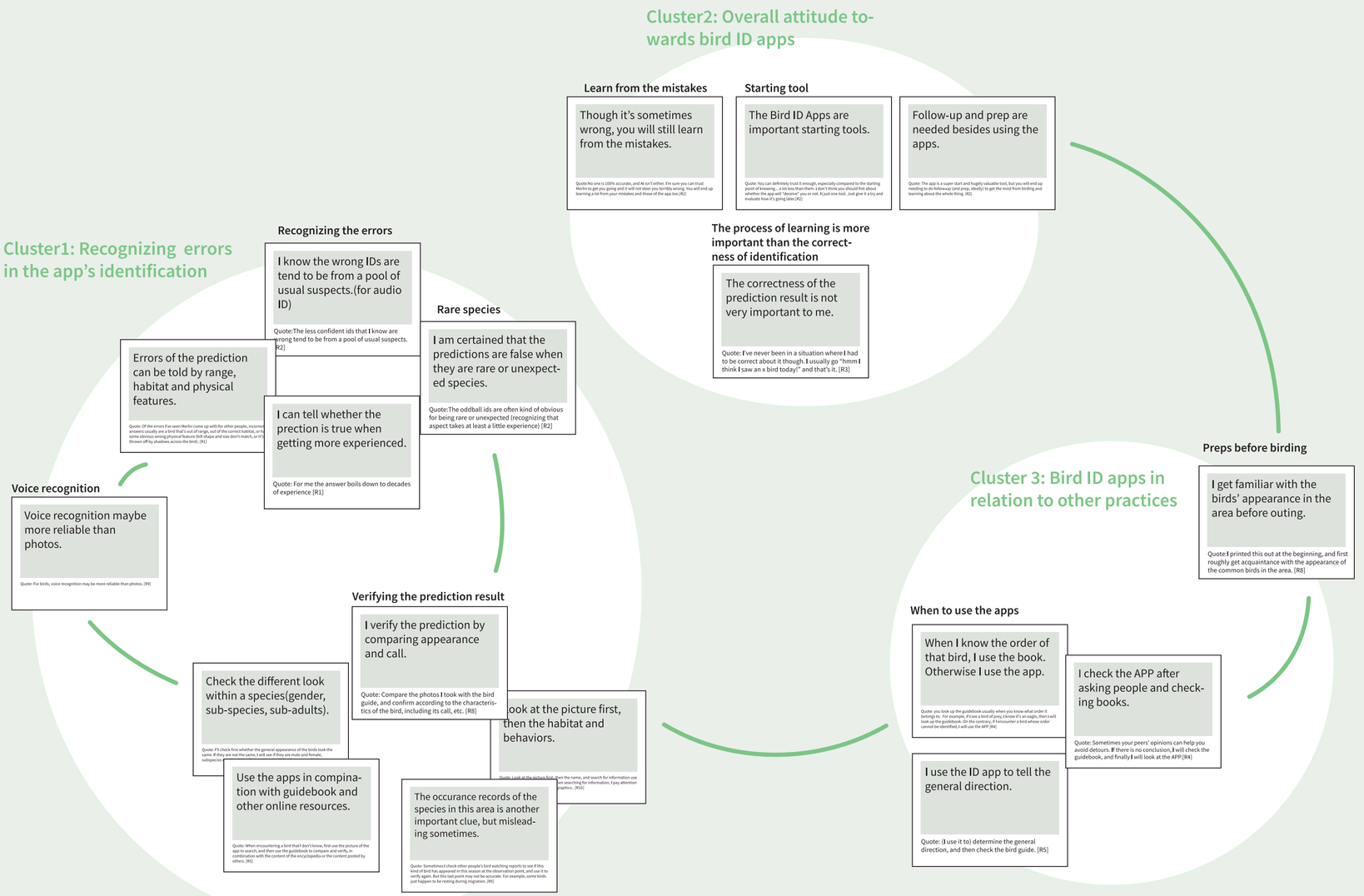


Figure 3-3. Clusterings in statement card analysis

3.4 Findings

3.4.1 General learning practices of birders

The three interviewees have a lot in common. They all have birding expertise beyond average and started birding in a birding club. They have tried, but none of them were active users of a bird ID app.

The findings from the **interviews** focused on the participants' general birding behaviors. In the interviews, they shared mainly about their activities and learning path apart from the bird ID apps.

> **RQ1**: What are the tools the birders currently use to learn about birds?
>
> a. What knowledge do they learn with these tools?

1. Lectures and knowledgeable birders helped them get started.

Stories of these 3 interviewees who got started were similar. They all picked up birding as a hobby by joining a birding club at school, where they took lectures and went birding with other birders to learn about birds.

In the lectures for beginners in the clubs, they were taught basic knowledge of bird taxonomy, anatomy, and tips for watching birds outside. And during the bird-watching tour, birders who are more experienced would teach them the birds' appearance, behaviors, habitats, and so on, which deepened the knowledge they learned.

> *"Firstly, the lectures and the guidance from experienced bird watchers helped build the system (of learning about birds), and that's the foundation of everything." - P1*

2. Bird books are the most reliable resource.

Aside from the lectures and skilled birders who got them started, they found the bird books to be the most helpful. During the interviews, they introduced to the researcher different kinds of bird books, either with illustrations or photos, of different regions, solely on the appearances or the behaviors of birds.

> *"Field guides are most helpful to me in learning. ...I prefer illustrated bird books to photo ones because it shows features that are more constant for certain species."-P2*

3. Among bird apps, they like the birding record tools and sound ID features.

When talking about bird apps, all of them found the birding record app (eBird) extremely useful, where they check in advance what birds are there around certain locations and upload their findings after a birding.

> *"I check the Hotspot on eBird before birding to see what species are around."-P1*

And the feature of ID through sound recordings was found useful by them.

> *"I found the sound identification feature very useful to me when it's hard to capture a clear photo. The Bird ID Master and Xeno both have such function."-P1*

> **RQ2**: What do they find most helpful/challenging in their process of learning to identify birds?

4. They don't rely too much on bird ID apps but discuss with other birders instead.

Because they are all members of a birding community and have experienced birders they trust around them, and because they normally go birding with others, they would ask their peer birders if they came across any unknown birds.

> *(Researcher: Have you tried any (image-based) bird ID apps?)*
>
> *"I have tried the Bird ID Master but not too much. Most of the time I use it as a digital dictionary to look up entries."-P2*
>
> *"I don't need a bird ID app when I'm in China because I'm familiar with most of the common birds here. I would probably use bird ID apps when I go to the U.S cause I don't recognize the birds in North America." -P3*

5. Subtle differences in appearance and difficulty in seeing hidden birds are the main challenges for identification.

When asked about the challenges in telling birds apart, they mentioned the difficulty of recognizing subtle differences in birds' appearance, and the difficulty in seeing a bird hidden in the woods.

> *"The subtle differences in feather color between some birds are hard to distinguish. And there are birds that couldn't be told apart merely with their appearance, for example, the Kamchatka, Japan and Arctic Leaf Warblers"-P2*
>
> *"When doing bird watching outside, birds are often hidden behind woods, it's hard to see and can only be identified by their calls."-P2*

3.4.2 Current practices of using bird apps

> **RQ3**: What do the birders currently use bird apps for?

While the **mini-survey** focused specifically on people's experience with bird apps (especially ID apps), we gained insights mainly on their opinions on those apps.

6. Bird ID apps are good starting tools for beginners even if they make mistakes.

As one of the participants (R2) put it, the ID apps will be good starting tools for the beginner birders, *"especially when you know a lot less than the apps ."(-R2)*

The beginners either trust the ID apps enough or don't care that much about the accuracy.

> *"I've never been in a situation where I had to be correct about it though. I usually go "hmm I think I saw an x bird today!" and that's it. "-R3*
>
> *"You can definitely trust it enough, especially compared to the starting point of knowing... a lot less than them. I don't think you should fret about whether the app will "deceive" you or not. It is just one tool. Just give it a try and evaluate how it's going later."-R2*

7. The learning happens in the process of exploring, even from the mistakes of the ID apps.

And they held the opinion that it is not the correctness of identification itself that matters the most, but the whole learning process. The point is people can learn a lot more beyond what a particular bird they spotted is, by doing preparation and follow-ups.

> *"No one is 100% accurate, and AI isn't either. I'm sure you can trust Merlin to get you going and it will not steer you terribly wrong. You will end up learning a lot from your mistakes and those of the app too."-R2*

> *"The app is a super start and hugely valuable tool, but you will end up needing to do follow-up (and prep, ideally) to get the most from birding and learning about the whole thing. "-R2*

8. With expertise growing, birders are able to tell the wrong predictions from the correct ones.

It gets easy for the birders to tell the wrong predictions as they gain more experience. Birders with at least a little experience can tell the false predictions by the physical features, habitats, and the rarity of the predicted species.

> *"Incorrect answers usually are a bird that's out of range, out of the correct habitat, or has some obvious wrong physical feature (bill shape and size don't match, or it's thrown off by shadows across the bird)." -R1*

> *"The less confident ids that I know are wrong tend to be from a pool of usual suspects. "-R2*

9.Birders use bird ID apps usually in combination with other resources.

5 out of 10 responses indicated that they use the identification apps in combination with other learning resources.

The birders currently use bird ID apps in combination with other practices to tell birds apart, depending on what situation they are facing, for example, do they want to know the specific species when they already have a ID direction in mind, or do they want to know roughly the order of the bird they spotted.

One of the replies indicates under what circumstances she/he will choose to adopt the bird ID apps instead of other tools like bird books.

> *"You look up the guidebook usually when you know what order it belongs to. For example, if I see a bird of prey, I know it's an eagle, then I will look up the guidebook. On the contrary, if I encounter a bird whose order cannot be identified, I will use the APP." -R4*

> *"(I use the ID apps to) determine the general direction, and then check the bird guide. "-R5*

Throughout a birding, they collect information before bird watching outside. During their outing, they capture pictures and sound of birds they are not sure about to recognize. After the outing they do follow-ups learning for those birds.

> *"The other resource is the official website of the city or the local bird group, which will publish the local bird pictures, such as the commonly seen birds in the xxx area. I printed this out at the beginning, and first roughly got acquainted with the appearance of the common birds in the area."-R8*

3.5 Conclusions

The qualitative research presented in this chapter provided insight into possible reasons why birders use or reject (photo-based) bird ID apps, as well as where these applications might fit into their learning process.

In the 3 semi-structured interviews, we learnt stories of 3 experienced birders, and received insights into how they use different learning materials and tools in their learning process, as well as what type of knowledge they learn from different channels.

In this way, we knew what sort of knowledge that birders need and want to learn, compared to what can be offered by bird ID apps or similar products.

On the one hand, learning to identify bird species is an important element of a birder's education, but it isn't the only aspect. Birders study birds not just so they can tell them apart, but also so they may discover intriguing things about their behaviour and so on.

Birders, on the other hand, identify bird species based on a variety of factors such as appearance, habitat, habits, flying, and sounds. And the information offered by a photo-based bird ID app can only help them learn to recognize birds based on their visual characteristics.

Thus for our project, with data enabled by photo classification models, we want to make it clear that we aimed only at the purple bit of the Venn diagram shown below, which means teaching birders to identify birds by their appearance (figure 3-4) .

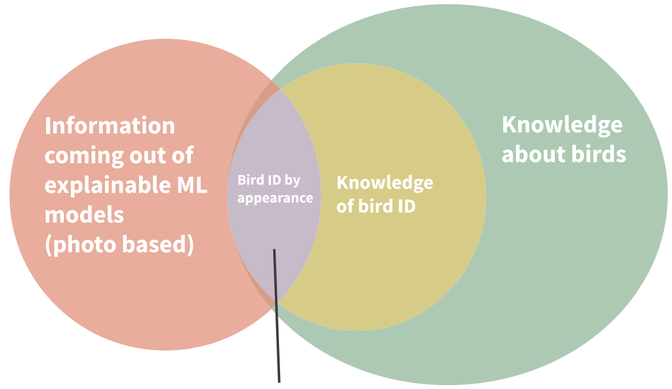


Figure 3-4. Venn diagram of bird knowledge that are supported by ML bird ID explainability

In mini surveys, we have learned how users currently use bird ID apps in combination with other tools to learn bird knowledge. Based on their statements, we drew a journey map of birders' typical practices of using bird ID apps (figure 3-5). It shows how the birders use the bird apps before, during, and after a bird watching practice, along with types of bird knowledge involved and their main struggles at each stage.

Identifying a bird with bird ID apps

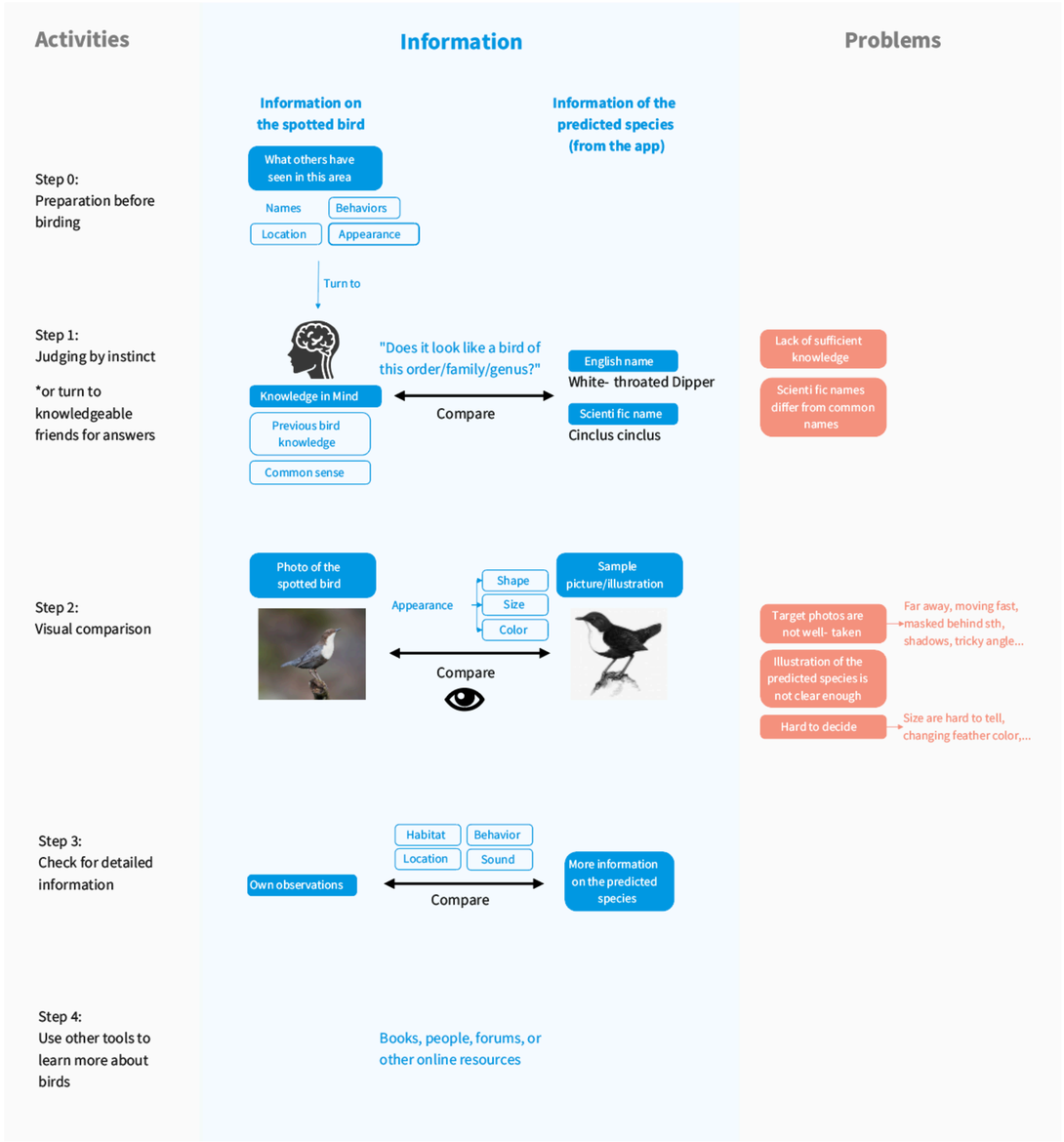


Figure 3-5. Journey map of birders' typical practices of using bird ID apps to identify birds

Summing up Chapter 3

The purpose of this phase is to gain a basic grasp of the birding context and evaluate how our proposed product may fit into people's existing learning paths. We learned how experienced birders came into the field of birding, their current practices, and useful tools and problems during their learning of bird identification by conducting three interviews with experienced birders. We learned how birders now use bird ID apps, what they value, and care about among the information provided through a mini-survey on social media, for which we received 10 responses.

The insights gained will serve as the basis of the set-up of the following quantitative research.

Takeaways

- Experienced birders interviewed learned birding through lectures, knowledgeable friends and field guides, which they think built them a solid knowledge foundation.
- The bird ID apps will be especially useful for beginners, who have much less knowledge than the apps.
- Though the predictions made by bird ID apps are not always correct, the users believe they can still learn something from it. And they value the learning process of bird knowledge over the correctness of the prediction result.
- Compared to using a bird ID app to identify a bird, more learning happens in their preparation and follow-ups, during their discussions with others, or looking into bird books.

CHAPTER 4.
THE ONLINE SURVEY

Main RQ: What are opportunities and challenges for different levels of birders to adopt bird ID apps as their learning tools?

With the insights from qualitative research, this chapter goes through the online survey carried out among birders.

In the last chapter, we noted some variations between novice and experienced birders. The goal of the online survey is to confirm the results of the qualitative research with a larger group of birders.

Furthermore, we want to understand what different levels of birders already use bird ID apps for, as well as their motivations, aims, and problems when using the apps, in order to decide who we are designing for and to create distinct personas for them.

4.1 Background

In our earlier qualitative study of birders' activities, we discovered that beginners and experienced birders have distinct attitudes on using bird ID apps. Experienced birders, for example, are less interested in bird ID apps, although some beginners find them to be quite useful learning tools. And we wanted to know if these findings could be generalized.

The goal of this phase was to learn about birders' habits and attitudes about bird ID apps in a quantitative approach, to see what opportunities there are for bird ID apps among different levels of birders. As a result, we have a clearer image of who we're designing for, as well as what they expect and need.

Research Questions

RQ4: How important are the bird ID apps in the learning process of different levels of birders?

RQ5: What are the different levels of birders' motivation for using bird ID apps?

4.2 Method

4.2.1 Questions

The online survey was set up to understand the birders' experience and learning preferences in relation to their different levels of expertise.

The survey consists of 4 parts:

1. Introduction and informed consent

2. Demographic questions

(age, professional relationship with ornithology or machine learning)

3. Questions on birding experience

4. Questions on experience with ID apps

After a brief introduction and informed consent information page, the first part was demographic questions about their age, and whether they have a professional relationship with ornithology or machine learning.

The second part was the questions on their birding experience in general, including what level they think they are at in telling birds apart, and what activities and resources have helped them in learning.

And the third part was the questions related to their experience with bird apps, including their motivations, behaviors and struggles while using them.

The complete questions set-up is documented in Appendix D.

4.2.2 Participant selection

This online survey was aimed at birders of all levels, including individuals who are interested in birding but have little experience. The online questionnaires were posted to birdwatching interest groups on social media platforms such as Reddit, Douban (Chinese version Reddit) and Wechat to reach the target audience.

4.3 Procedure

4.3.1 Participants demographic

49 replies were collected in total, about 44.90% participants fell in the age group 16-25, 30.61% of them in 26-35 (figure 4-1). When it comes to expertise in telling birds apart, more than 70% of them have put themselves in entrance or intermediate level, while less than 30% are in advanced or expert level (figure 4-2).

The second half of the questionnaire was about their experience with bird ID apps, only those who have or may have tried bird apps will be shown with the questions. 36 out of 49 participants indicated that they have or may have tried bird ID apps, and answered the questions on bird ID apps (figure 4-3).

Q1- What's your age?

Answer	%	Count
0-15	2.04%	1
16-25	44.90%	22
26-35	30.61%	15
36-45	6.12%	3
46-55	16.33%	8
>55	0.00%	0
Total	100%	49

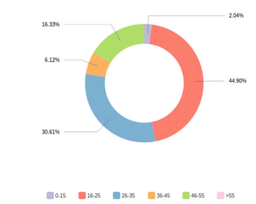


Figure 4-1. Responses for Q1

Q8- What level you are at in telling birds apart, compared to people around you?

Answer	%	Count
Entrance-level	24.49%	12
Intermediate-level	48.98%	24
Advanced-level	20.41%	10
Expert-level	6.12%	3
Total	100%	49

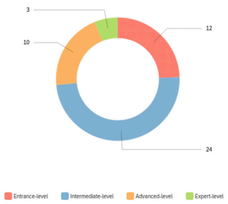
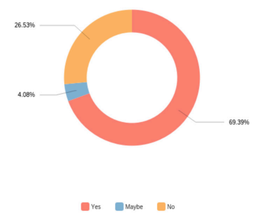


Figure 4-2. Responses for Q8

Q14 - Have you tried any bird ID apps?



#	Answer	%	Count
1	Yes	69.39%	34
2	Maybe	4.08%	2
3	No	26.53%	13
	Total	100%	49

Result broken down by expertise

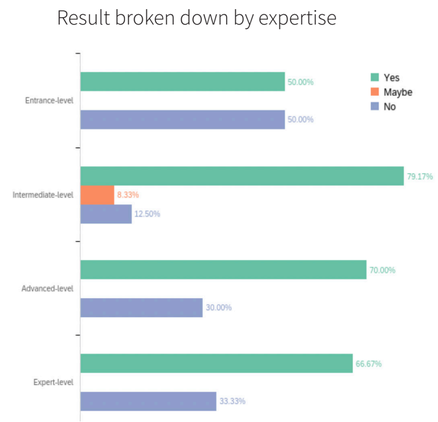


Figure 4-3. Responses for Q4

4.3.2 Hierarchical analysis

The survey results were broken down by the participants' expertise to compare the differences in practices of different levels of birders.

The data used to compare were: the importance of different learning resources to them, their motivations and purposes for using the bird apps. The breakdown results are shown in figure 4-4.

Q12 - What references do you usually use for learning about birds? And please rate them based how important they are for your learning. (0=not important/not used, 5=very important)

#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
4	Books or fieldguides of birds	0.00	5.00	4.20	1.32	1.75	49
3	Knowledgeable friends/experts around me	0.00	5.00	4.00	1.32	1.76	49
8	Bird ID apps (eg. Merlin Bird ID, Bird ID Master)	0.00	5.00	3.33	1.66	2.76	48
7	Birding records of other birders (eg. eBird)	0.00	5.00	3.08	1.72	2.97	49
6	Other online resources	0.00	5.00	2.94	1.64	2.68	48
5	Discussions in online forums / interest groups	0.00	5.00	2.92	1.64	2.70	48
1	Lectures of bird watching	0.00	5.00	2.19	1.77	3.13	47
2	Video and DVD guides	0.00	5.00	2.13	1.72	2.96	45

Result broken down by expertise

Entrance-level

#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
3	Knowledgeable friends/ experts around me	0.00	5.00	3.42	1.98	3.91	12
4	Books or fieldguides of birds	0.00	5.00	3.25	1.96	3.85	12
6	Other online resources	0.00	5.00	3.08	1.71	2.91	12
8	Bird ID apps (eg. Merlin Bird ID, Bird ID Master)	0.00	5.00	2.92	1.80	3.24	12
5	Discussions in online forums / interest groups	0.00	5.00	2.75	1.96	3.85	12
7	Birding records of other birders (eg. eBird)	0.00	5.00	2.58	1.55	2.41	12
2	Video and DVD guides	0.00	5.00	2.58	1.98	3.91	12
1	Lectures of bird watching	0.00	5.00	2.25	1.88	3.52	12

Intermediate-level

#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
4	Books or fieldguides of birds	2.00	5.00	4.38	0.95	0.90	24
3	Knowledgeable friends/ experts around me	3.00	5.00	4.33	0.80	0.64	24
8	Bird ID apps (eg. Merlin Bird ID, Bird ID Master)	0.00	5.00	3.88	1.62	2.61	24
7	Birding records of other birders (eg. eBird)	0.00	5.00	3.08	1.80	3.24	24
5	Discussions in online forums / interest groups	0.00	5.00	3.00	1.47	2.17	24
6	Other online resources	0.00	5.00	2.79	1.63	2.66	24
1	Lectures of bird watching	0.00	5.00	2.38	1.65	2.73	24
2	Video and DVD guides	0.00	5.00	2.04	1.55	2.39	23

Advanced-level

#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
4	Books or fieldguides of birds	4.00	5.00	4.80	0.40	0.16	10
7	Birding records of other birders (eg. eBird)	2.00	5.00	4.00	1.00	1.00	10
3	Knowledgeable friends/ experts around me	1.00	5.00	3.60	1.11	1.24	10
5	Discussions in online forums / interest groups	0.00	5.00	3.22	1.69	2.84	9
8	Bird ID apps (eg. Merlin Bird ID, Bird ID Master)	1.00	4.00	3.00	0.94	0.89	9
6	Other online resources	0.00	4.00	2.67	1.49	2.22	9
1	Lectures of bird watching	0.00	4.00	1.63	1.65	2.73	8
2	Video and DVD guides	0.00	4.00	1.14	1.36	1.84	7

Expert-level

#	Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
3	Knowledgeable friends/ experts around me	5.00	5.00	5.00	0.00	0.00	3
4	Books or fieldguides of birds	4.00	5.00	4.67	0.47	0.22	3
6	Other online resources	3.00	5.00	4.33	0.94	0.89	3
2	Video and DVD guides	2.00	5.00	3.33	1.25	1.56	3
7	Birding records of other birders (eg. eBird)	0.00	5.00	2.00	2.16	4.67	3
5	Discussions in online forums / interest groups	1.00	3.00	2.00	0.82	0.67	3
1	Lectures of bird watching	0.00	5.00	2.00	2.16	4.67	3
8	Bird ID apps (eg. Merlin Bird ID, Bird ID Master)	0.00	3.00	1.67	1.25	1.56	3

Figure 4-4. Responses for Q12

4.4 Findings

RQ4: How important are the bird ID apps in the learning process of different levels of birders?

1. Bird ID apps are important learning resource for beginners, next to bird books and their knowledgeable friends

In general, bird books are the most useful learning resources for bird hobbyists, which is in line with the result from the previous stage. And the importance of bird apps ranked after bird books and their knowledgeable friends (figure 4-4).

Breaking the result down by expertise (figure 4-4), it is found that as birders get more advanced in telling birds apart, the role bird ID apps play in their learning gets less important. This could be explained by the fact that the width of knowledge provided by bird apps is limited thus couldn't meet the needs of experienced birders.

One response wrote:

> *"At first, I used it mainly as a way to ID unknown birds. But after my 4th or 5th outing, I really don't use bird ID apps. However, I do prep in advance. I check local sightings of birds and study what is in my area." (Reply for Q18: How do you use the bird ID apps?)*

When comparing the scores rated by entrance-level and intermediate-level hobbyists, the importance score of the bird apps rated by entrance-level hobbyists (mean=2.92) is lower than that rated by intermediate-level hobbyists (mean=3.88). This makes sense, despite the previous finding, because half of the newbie enthusiasts haven't tried any bird applications at all (figure 4-3).

One reply from entrance-level wrote:

> *"Mostly I don't look them up on purpose, I just let it be. Sometimes when I come across some super pretty birds, I would take a photo of them and search." (Reply for Q13:What do you usually do to tell birds apart?)*

RQ5: What are the different levels of birders' motivation for using bird ID apps?

2. People use bird apps mostly because it's easy to access

The result shows that the convenience of use is a crucial influencing factor for people's decision of adopting bird apps (figure 4-5).

For those who have tried bird apps, they were asked what their motivation for using bird apps is. And the top two reasons are because of its convenience to reach and its convenience of finding the answer.

For those who haven't used bird apps, one of the replies says:

> *"(I don't use them because) Everytime I see a bird I am outside, and downloading an app outside will cost me data." (Comment from one intermediate-level participant)*

3. People use bird ID apps as digital bird books, most beginners consider it as their main learning tool

When asked what they use bird apps for, "I use it as a digital bird book to look up entries" was the top answer (figure 4-6). This means the participants don't

use only the photo ID feature but also use the bird apps to gain in-depth bird knowledge.

Breaking down, many entrance-level birders (66.67%) use bird apps as the main tool to learn about bird identification, the percentage is 28.57% among intermediates and drops to 0 among higher-level birders.

4. Bird ID apps sometimes went wrong and the beginners couldn't tell

Of all the participants, most of them have encountered errors in the prediction. Among birders from the entrance and intermediate levels, many of them are unsure about the correctness of the prediction result (figure 4-7).

Q17 - Why do you use the bird ID apps? (multi-choice)

Answer	%	Count
Because it's convenient to reach compared to a bird guide	83.33%	30
I can find the answer more quickly than looking into a bird guide	63.89%	23
Because I can show it to others easily	30.56%	11
Because I'm curious whether it could identify correctly	19.44%	7
I'm not a fan of bird ID apps	2.78%	1

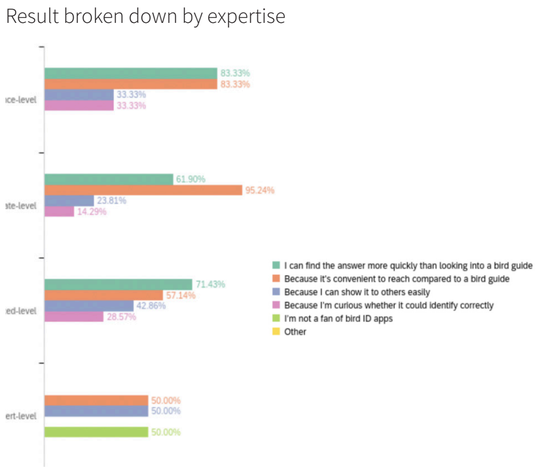
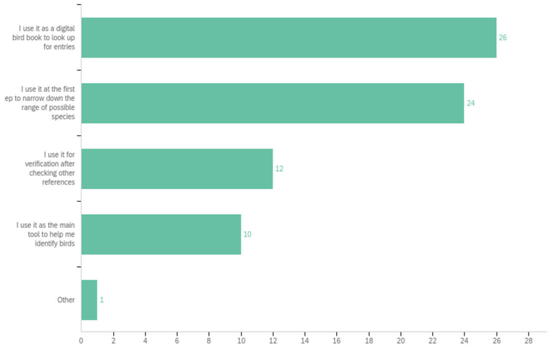


Figure 4-5. Responses for Q17

Q18 - How do you use the bird ID apps? (multi-choice)



Result broken down by expertise

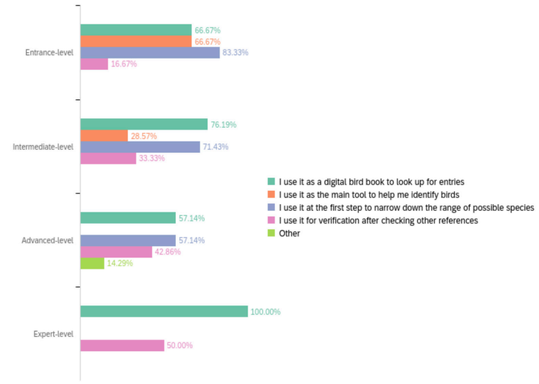


Figure 4-6. Responses for Q18

Q19 - Have you encountered false predictions provided by the ID apps?

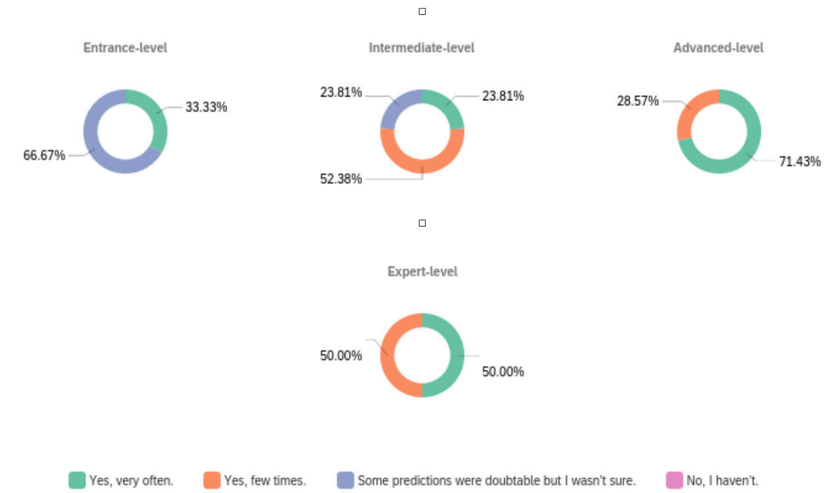


Figure 4-7. Responses for Q19

4.5 Conclusions

Based on the analysis of the online survey result, personas were made for different levels of bird hobbyists (figure 4-8).

The entrance-level respondents in the survey were furtherly divided into complete novice and beginners, the former of which has nearly zero birding experience while the latter has at least a little, as we acknowledged there were quite some differences between them. And as we don't have enough samples for advanced or expert birders from this online survey, these two kinds of birders were put together.

The result of the online survey shows many of the entrance-level birders tend to use the bird apps as their main tools for learning about birds. This might be because of its convenience compared to other learning resources such as bird books. And intermediate and more advanced birders would use the bird apps in combination with other tools for more in-depth bird knowledge, this might be explained by the fact that they are more sensitive to accuracy of the information acquired.

Based on the users' persona and the analysis above, the complete novice and the beginners are most likely to be the group of people that we will design for as they mostly rely on the bird identification apps to learn bird knowledge. Which means they are likely to adopt our product for the purpose of learning. And the intermediate-level birders, with 28.57% of whom would use bird ID apps as their main learning tool, could also be considered as the potential users whom we would design for.

We wish to recruit as many users as possible for data gathering purposes as a citizen science project. So it's good to find out that people with little expertise may be interested in using it.

Advanced birders, on the other hand, may be motivated to use our product for other reasons, such as a sense of accomplishment, even if they don't utilize the bird ID applications as their primary learning tool. These assumptions will be further explored by presenting them with experiential prototypes in the following stage.

Complete Novice

Descriptions

- Have zero bird knowledge before but would like to learn some out of curiosity.

Current Behaviors

- Enjoy watching birds casually in daily life, but never bother knowing exactly what birds they are.
- Haven't used any bird apps or owned any bird books. Sometimes search online about birds.

Needs

- Want to learn about birds in an interesting way without putting too much effort or time in it.

Intermediate

Descriptions

- People whose bird knowledge are beyond average. Take birding as a serious hobby and have interest in learning more in-depth knowledge of ornithology.

Current Behaviors

- Go bird-watching on pupose and on a regular basis. Fieldguides are the most helpful resource for learning.

Needs

- Want to acquire systematic bird knowledge even if though it takes time. The accuracy of ID is important to them.

Beginner

Descriptions

- People who doesn't have much bird knowledge but decide to develop birding as a hobby.

Current Behaviors

- Look up tutorials online and read bird books.
- Use mostly the photo ID feature of bird apps. But find it difficult sometimes to know whether the prediction is trustworthy.

Needs

- Want to learn to identify common birds in the neighbourhood in a quick and easy way.
- Would like to learn something beyond only the appearance of birds.
- Nevertheless the correctness of ID doesn't bother them that much as they are doing it just for fun.

Advanced/Expert

Descriptions

- People who are familiar with most regular birds around their places. And know a lot about birds' behaviors, migration, etc.

Current Behaviors

- Check bird books or discuss with friends for most problems they have.
- Don't need bird ID apps, unless when they are going to a new place. They are with lots of bird knowledge equipped and am able to tell the correctness of the predictions of ID apps.

Needs

(Unclear due to limited samples size)

Figure 4-8. Persona of different level's birding hobbyists

Summing Up Chapter 3

In this stage, I sent out an online survey to birding hobbyists with different levels of birding expertise, to know the role that the bird apps play in their learning of bird knowledge. The results of the online survey reveal great differences in their preferred learning tools, and in their motivations for using bird apps. In other words, the entrance-level birders care more about the convenience and fun in their learning tour, while more advanced birders care more about the professionality of the content and will thus choose more professional learning tools. Starting from their different preferences, I chose to design mainly for the complete novice and beginner hobbyists.

Takeaways

- Compared to birders of all the other levels, entrance-level birders are more likely to adopt bird ID apps as their main learning tools for bird knowledge.
- It is hard for entrance-level birders to tell the wrong predictions from the right ones.
- People use bird ID apps rather than bird books because they value convenience in learning.
- The complete novices and beginners in birding are chosen as the main target users of our product. The intermediates are the potential users.

CHAPTER 5.
THE EXPLANATION PROTOTYPES

Main RQ: What are the design opportunities for explainable AI models in the birding community?

Although ML explainability has been widely utilized to promote human-computer trust, our previous research has found that our target users are unconcerned about trust issues. What is the most important aspect of ML explainability for people? Is our product still required to provide justification in the context of birding? We want to find out what they want and need from the explanations in ML models at this point after identifying the target user group in the previous chapter.

Firstly, based on the taxonomy of the machine learning explanation, three quick explanation mock-ups were built for tests.

Then, we sought to connect the needs of the end-users and the capabilities of the explanation methods, by presenting interfaces of the prototypes to the target users and let them rate how those prototypes achieve their needs in different aspects.

5.1 Background

In the previous chapter, we learned which group of birders is most likely to use bird ID apps, what they use them for, and their challenges.

In Chapter 1, we conceived the framework of an explainability method that can collect annotations from end-users while benefiting themselves in some ways as their motive of using.

Traditionally, machine learning explanations have always aimed to increase human-computer trust (Gunning et al., 2019). However, during the earlier stages of the research (Chapter3, Chapter4), trust issues did not appear to be a major concern for our target users; instead, they seemed to be more concerned with the learning process.

To find out more about this, we developed three different explanation prototypes as sensitizers, and we drew assumptions based on the explanation goals summarized by Nothdurft et al.(2013).

Goal of Explanation	Description
Justification	Explain the motives of the answer?
Transparency	How was the systems answer reached?
Relevance	Why is the answer a relevant answer?
Conceptualization	Clarify the meaning of concepts
Learning	Learn something about the domain

Figure 5-1. The different goals an explanation can pursue (Nothdurft et al., 2013)

Of all the five goals (functionalities), we assumed the goal of justification, transparency, and learning will be the most relevant ones to users in our context, and use them as part of the metrics of the evaluation to know how well do the prototypes fulfill these goals.

> **RQ6:** Which explanation goals are most valued by the target users?

We'd also like to know the following things to help in follow-up development:

> **RQ7:** Do the explanations of the bird species classification in reaching that goal?
>
> a. Which properties of the prototypes help in reaching the goal?
>
> **RQ8**: What else do they expect from an educational bird app with explanations?
>
> **RQ9:** What could motivate the target users to take part in the annotation process?

5.2 The Test Prototypes

5.2.1 Design space

Though we've opted to use the SECA framework to improve explainability, we haven't settled on how the explanatory information will be delivered. For ideation, we will look at the taxonomies of machine learning explanation to find properties. While not all of the variants can be realized in a SECA framework, we pick out only those properties that fall into the capability scope of the SECA framework.

As introduced in Chapter 2, there are multiple ways to categorize machine learning explainability approaches. For example, taxonomies based on functional requirements of the approach, or that based on operational requirements. (Sokol & Flach, 2020)

However, the ideation of this stage focuses only on the following two dimensions of the explanation approaches:

Explanation target (functional): What types of information are to be explained?

Explanation Medium (operational): How will the information be presented to the end-users?

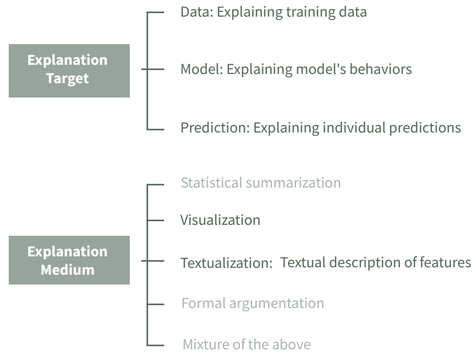


Figure 5-2. The explanation variants under the property of Explanation Target and Explanation Medium

The data, model, and predictions are all part of the machine learning process. The SECA explains the overall behaviors of the model rather than individual predictions since it is a global explainability method rather than a local one. (Balayn, 2021) This left the data and the model as options for the first questions (explain target). Explaining the data refers to explaining the training data, whereas explaining the model refers to the model's behaviors.

The output of SECA is a textual description of the saliency maps' highlighted areas. As a result, we believe that such explanations will most likely be presented in the second dimension (explain medium) as visuals, textualization, or a combination of the two.

5.2.2 Ideation of prototypes

In the ideation session, ingredients from the two dimensions were combined together to be several different possible design directions. With the knowledge of the different possibilities there are, different design strategies were mapped to the goals or struggles of the target users (figure 5-3) .

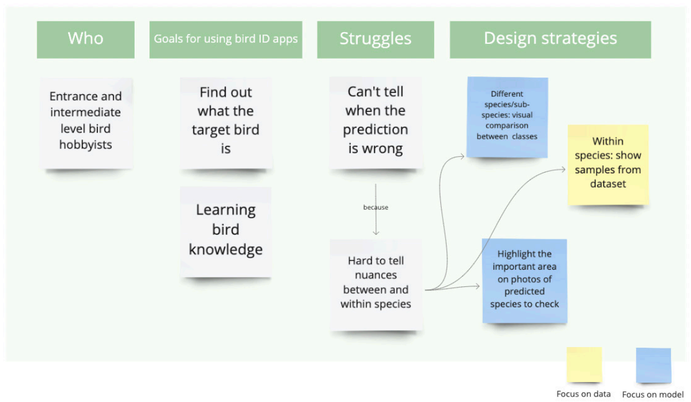


Figure 5-3. Design strategies mapped to the users' struggles

After this, these different design strategies were converged and sketched out to be the primary prototype ideas for tests (figure 5-4). In practice, the concept of comparison was broken down into a prototype of “feature description” and one of “result comparison”, in order to study the property of textual description and comparison separately.

5.2.3 Three prototypes for test

The web prototype for the test contains a normal input and prediction page as start page and 3 pages of add-on explanation features (figure 5-5). The add-on features were separately named as “feature description”, “result comparison”, and “showing samples” (see Appendix F for the complete prototypes).

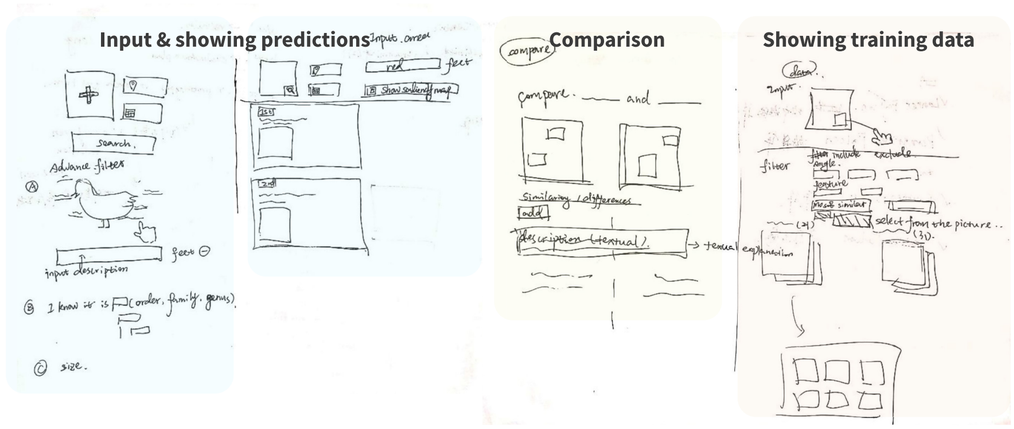


Figure 5-4. Sketches of the prototyping ideas

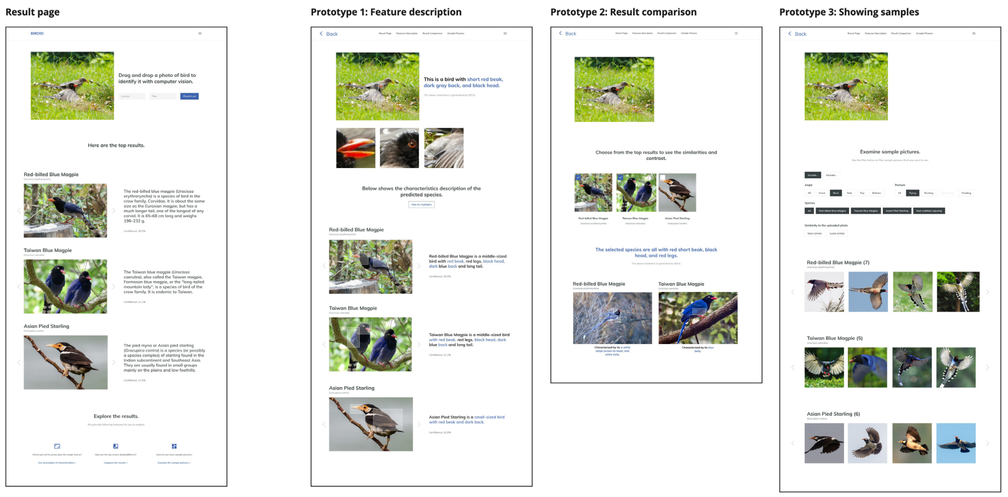


Figure 5-5. The prototype pages

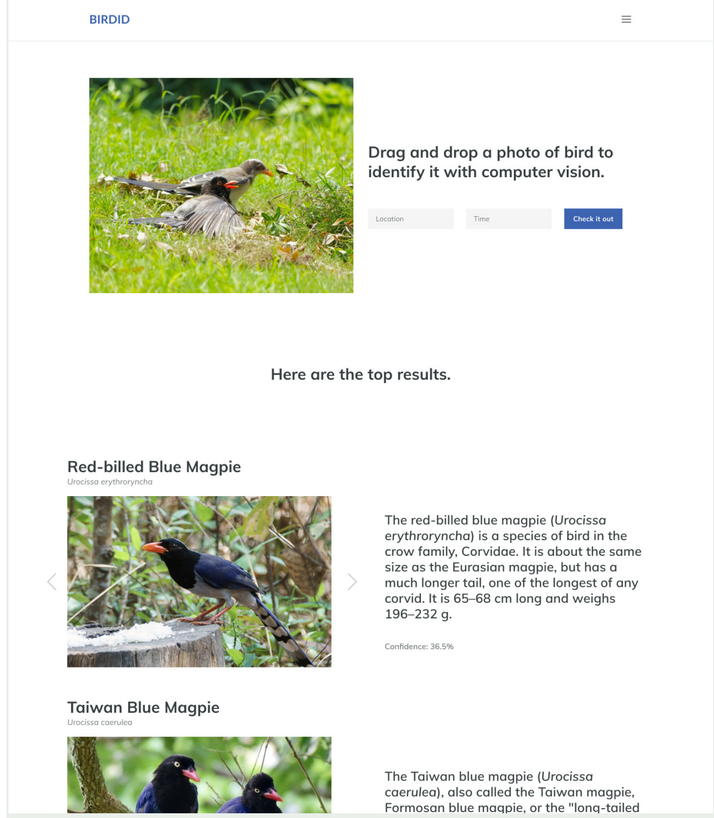


Figure 5-6. Normal prediction page

Normal prediction page (left)

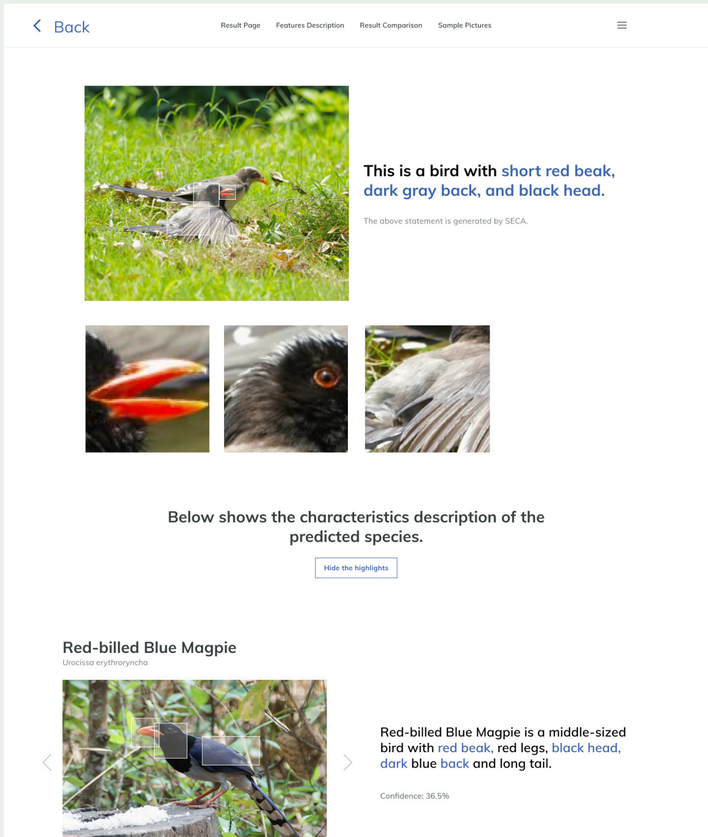
Description: A web page simply showing all the prediction results with a piece of brief information for each, is presented to the participants to provide a full experience of identifying birds with a web app. Participants were also asked to use this as a baseline, by rating how different add-on features add to the experience.

Prototype 1: Feature description (referred to as “Description”) (right)

Description: This prototype aims at showing the highlights of the salient characteristics recognized by the computer, in the form of displaying bounding boxes highlighting the identified features, along with the textual description of the highlighted features.

Properties: confidence value; explanation on the target photo; bounding boxes highlighting salient features; textual annotation of salient features; one sample photo for each species.

Figure 5-7. Feature Description function



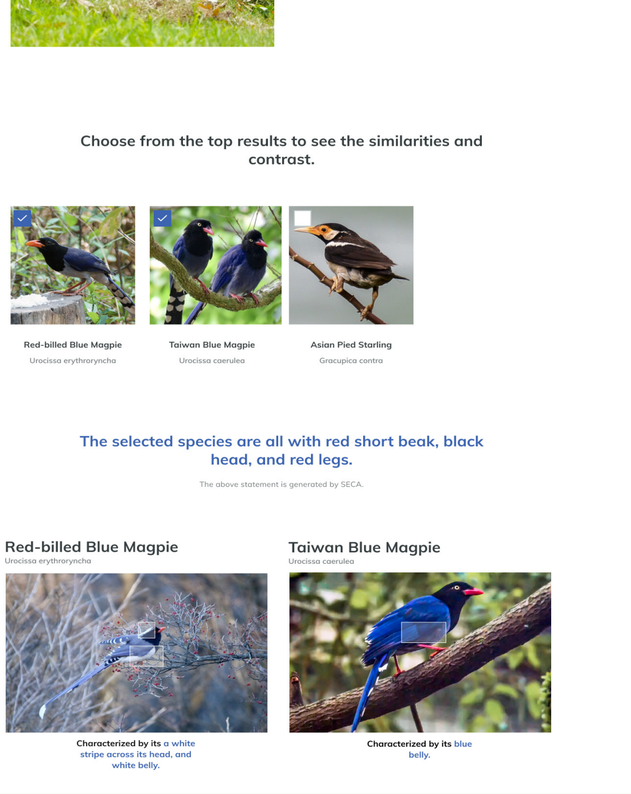


Figure 5-8. Result Comparison function

Prototype 2: Result comparison (referred to as “Comparison”) (left)

Description: This prototype provides users with the comparison of two prediction results. Users can select from the prediction list what they want to compare and see the similarity and contrast marked by bounding boxes on the photos, along with a textual description.

Properties: the contrast between prediction results; bounding boxes highlighting contrast features; textual annotation of contrast features; one sample photo for each species.

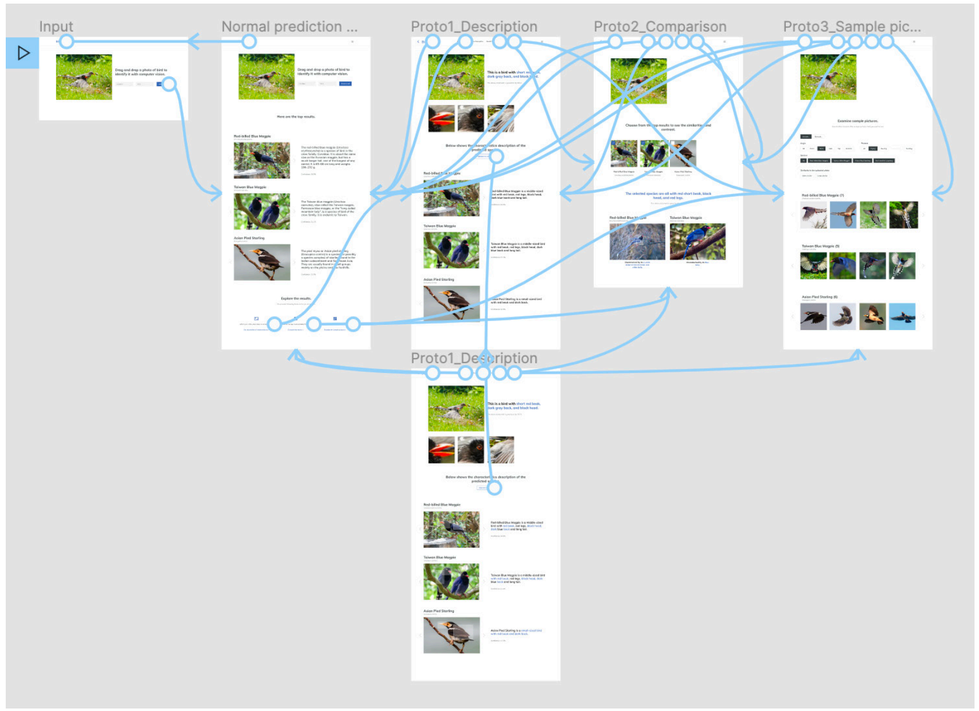


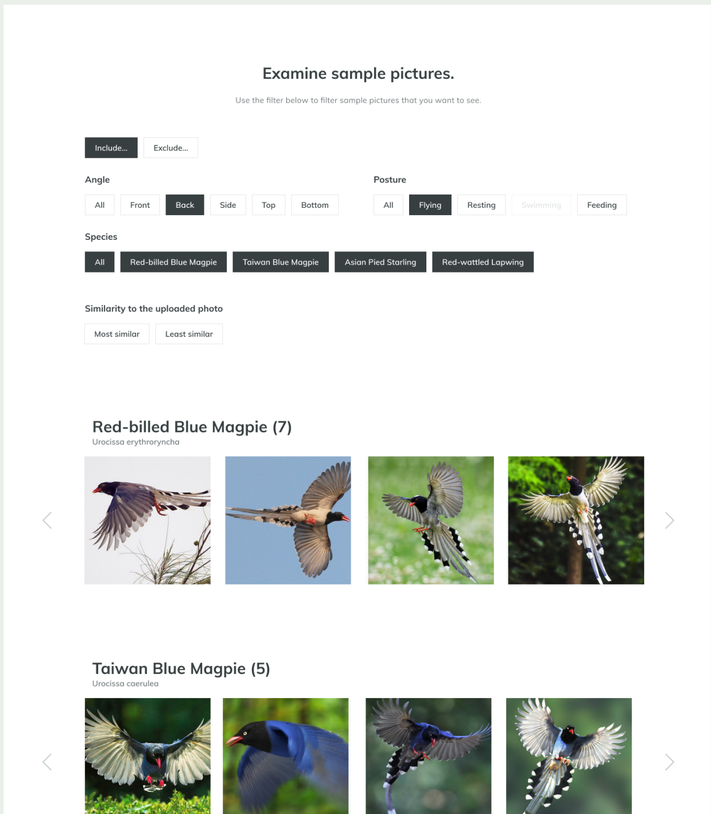
Figure 5-10. Prototype workflow in figma

Figure 5-9. Showing Samples function

Prototype 3: Showing samples (referred to as “Samples”) (right)

Description: This prototype aims at showing abundant sample pictures for users to examine. Filters of angles, postures, and similarities are provided for users to filter photos they need.

Properties: Providing sample pictures of predicted species; Filters for angles, postures, and similarity.



Minor interactions and animations(eg. showing and hiding bounding boxes) were developed in Figma for participants to navigate around these pages freely during the evaluation (figure 5-10).

5.3 Method

5.3.1 The evaluation metrics

We know that an explanation can pursue the goals of justification, transparency, relevance, conceptualization and enabling learning (Nothdurft et al., 2013).

And we make the guess that the justification, transparency, and learning function might be what our target users need in the chosen context. And this evaluation is to find out how our prototypes realize those goals and which of the goals suit their needs most.

We will figure out the first one by letting people rate, and the second one by asking them to rank the importance of these goals at the end.

So each of the following statements will be rated by the participants with a 5-scale likert.

> ***Transparency:*** *I think this function helps me understand why the predictions were made (and the reasons why certain predictions were false, if applicable)*
>
> ***Justification:*** *I think this function brings me more certainty on whether the prediction is true or false.*
>
> ***Learning:*** *I think this function enables me to learn more about birds.*

And the following two on prototypes' usability and people's preference.

Usability: *I think this function is easy to understand and easy to use.*

Like: *I like this function.*

5.3.2 The evaluation steps

The evaluation steps were as follow (see Appendix G for details):

1. Brief introduction to the project background, informed consent and the SECA framework.

2. Three opening questions:

-What level are you at telling birds apart?

-Have you tried any bird apps?

-Do you recognize this bird?

3. Exploring the prototypes: show the explanation prototype to the participants, ask them to explore all the prototypes freely for a couple of minutes (around 5mins).

4. Rating sessions: for every single prototype, there are ratings of 5 metrics, assessing how the prototype behaves in terms of its usability, transparency, justification, enabling learning, and how much it is liked. Each aspect will be rated with a 5-scale likert.

5. Generating new ideas: open questions were asked to gather their ideas on enriching the specific feature.

-Is there any other information you would like to see on the interface?

6. Needs for explanation: ask the participants to rank how different goals were important to them.

-What are the goals that you are pursuing when checking these prototypes?

-Which one is more important to you?

7. Attitude towards making annotations: check the possibility of involving the bird hobbyists in the annotation tasks.

-Would you be interested in contributing your own annotation when using this website? Why?

5.3.3 Participant selection

6 participants in total took part in this session. They participated in either one of the interviews, mini-survey, or the online survey, and were reached out by the researcher using the contact methods they left for this follow-up evaluation .

The participants were chosen by the researcher based on how experienced they are in birding. The recruitment guidelines were:

1) Recruit 4~5 birders from the beginners or intermediate levels as they were defined as the main target users.

2) Include 1~2 from advanced or expert level, to gain some advice from professional aspects.

The evaluation mainly focuses on birders from the entrance and intermediate level, as they are the target users of the conceived app (conclusion from Chapter 3). One birder from the advanced/expert level was also involved in the evaluation, to know professionals' advice on showing information.

5.4 Procedure

5.4.1 Data collection

All the evaluations were conducted via online Zoom meeting. Throughout the evaluation session, participants were asked to share their screen for screen recording. During and after each evaluation session, the researcher noted down participants' comments that were considered important in an evaluation template form.

After completing all 6 sessions, raw data of the answers and comments from the participants were documented in the Appendix H.

5.4.2 Data analysis

The analysis was conducted around the 4 research questions brought up in the background of this Chapter.

On the basis of the documented raw data, the researcher picked out quotes that were related to the research for analysis.

Specifically for RQ7, we clustered the feedback in relation to different properties of the prototypes and different explanation goals . In this way a clearer view was shown on how different properties could help in the participants' needs (figure 5-12, see Appendix I for complete analysis table).

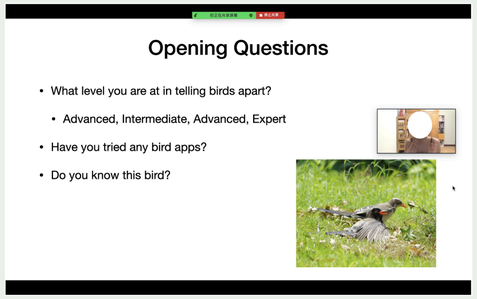


Figure 5-11. Captures of one of the evaluation sessions (via zoom), the opening session

Clusters of properties	Properties	Prototype NO.	Feedback	Relevant goal	Who holds a similar opinion?
Highlights on image and the textual description	Highlights on the target photo	1	💡"It will be more user-friendly if the highlights in the photo are presented in relation to the textual description."	(usability)	3(ABC)
	Description of the target photo	1	✅ "The description enables me to identify whether the prediction is trustworthy, or whether it fails because my uploaded photo is not good enough, for example, certain features not obvious."	✅Transparency, Justification	3(DEF)
	Highlights on the prediction photos	1	✅ "It taught me where to look at when seeing these birds."	✅ Learning	3(CDE)
		1	⛔"I think it shows exactly how the computer makes its prediction, but I doubt there's more to it for identifying birds. I want to see more detailed descriptions of their features instead of general ones."	⛔ Transparency, Learning	2(AB)
	Description of the predicted photos	1, 2	✅ "By providing info of features not shown in the uploaded photo, users are enabled to compare the info to what they have seen in the wild (but not captured by the camera)"	✅Transparency	4 (ACDF)
Comparison	Comparison between different prediction results	2	✅" This feature is useful for learning to distinguish birds"	✅ Learning	5 (ACDEF)
		2	⛔"I want to see the textual description more connected to the features highlighted in the photo."	(usability)	3 (ABC)
		2	⛔"I want the description to tell me directly the differences between them instead of the way it is now, preferably with a close-up of the features."	(usability)	2 (AC)
	Comparison between target and prediction	2,3	"I want to see the link between the target and predicted species, the fact that it's missing from the current prototype makes it hard to read." beginners, but very helpful for	(usability)	4 (ABCF)

Figure 5-12. Part of the feedback analysis process

5.4.3 Participants overview

In terms of their expertise in identifying birds, 3 out of the 6 participants put themselves in the entrance level in telling birds apart, 2 in the intermediate level and 1 in the advanced/expert level.

A letter between A and F was given to represent each participant.

In terms of their experience with bird identification apps, two participants (A and B) from the entrance level have never tried any bird apps. In addition to this, all the other 4 participants have tried at least one bird app, including eBird, Bird ID master, or digital bird books. They were also asked whether they recognize the bird in the test photo before examining the prototype. The participants demographic was shown in the following table (figure 5-13).

No.	Expertise	Tried bird ID apps?	Recognize the bird?
A	Complete novice	No	Not at all
B	Complete novice	Tried photo id apps	Not at all
C	Entrance	Tried some	Not at all
D	Intermediate	Tried some	Yes, but not sure
E	Intermediate	Tried some	Yes, but not sure
F	Advanced	Tried some	Yes, with certainty

Figure 5-13. Participants demographic

5.5 Findings

RQ6: Which explanation goals are most valued by the target users?

5.5.1 The main goal was to learn

When asked which of the explanation goals they valued most, the learning and justification (to know whether the prediction is true) seemed to be more important than the transparency (to know why computers make such predictions). Among the participants from intermediate or lower levels (target group), 3 out of 5 of them valued learning more than the other two goals, while the other 2 of them believed that other goals will eventually lead to learning.

"Knowing what the bird is is the most important goal. And then the justification so I trust the result."

- Participant A

"I care more about knowing what bird it is. And I want to learn something along the way. And I don't care about how the machine works(transparency)."

- Participant B

"When watching birds I first want to learn bird knowledge. Then I want to know what characteristics do the machines focus on, which lets me know what to pay attention to when watching birds.

Meanwhile, the accuracy of the prediction result is less important to me, I would rather ask around to get a good answer instead of relying on photo identification."

-Participant D

No.	Expertise	Ranking of goals
A	Complete novice	**Learning** > Justification > Transparency
B	Complete novice	**Learning**
C	Entrance	Justification
D	Intermediate	**Learning** > Transparency > Justification
E	Intermediate	Justification > Transparency > **Learning**
F	Advanced/Expert	Justification

Figure 5-14. How participants valued the explanation goals

Participants C and E, who placed learning after the other two goals, stated that this was because learning would naturally follow when the other goals were met.

"Justification is the most important goal. I don't trust the prediction result blindly, it has happened to every app I used when the predictions were doubtable, especially when the picture I uploaded was not very clear. "

-Participant C

"For me, the most important goal is to confirm which species it is, and knowing the mechanism of the models helps in reaching that goal. And the learning part will follow naturally after reaching those two goals, it's not that important but will add to it."

-Participant E

Participant F (advanced/expert) was the only one who didn't seem to mind learning much, owing to the fact that there wasn't much to learn for him in the information presented. As a result, he was primarily concerned with the accuracy of the prediction.

"I would feel more comfortable using this app because it provides me with clues to justify its predictions, so I don't feel confused or probably being deceived by it."

-Participant F

Based on the comments above, we can draw the conclusions that the birding community will find learning from the explanations most useful for them. Most of them don't care about the transparency goal. Some care about the justification goal, but for the reasons that they don't want to learn from false information.

RQ7: Do the explanations of the bird species classification in reaching that goal?

a. Which properties of the prototypes help in reaching the goal?

5.5.2 How different properties of the prototype help in learning

The following table (figure 5-15, 5-16, 5-17) shows the mean of the score that each prototype was rated on different aspects.

In the data analysis step (5.4.2), we have mapped all the participants' feedback to the corresponding properties and explanation goals (Appendix I).

Name of prototypes

	Description	Comparison	Samples
Usability	4.50	3.67	4.50
Transparency	4.60	3.40	3.10
Justification	4.00	3.83	3.33
Learning	2.50	3.83	3.83
Like	4.17	3.50	4.17

Figure 5-15. Average score of testing prototypes on different evaluation aspects

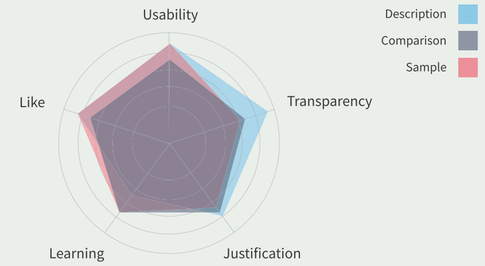
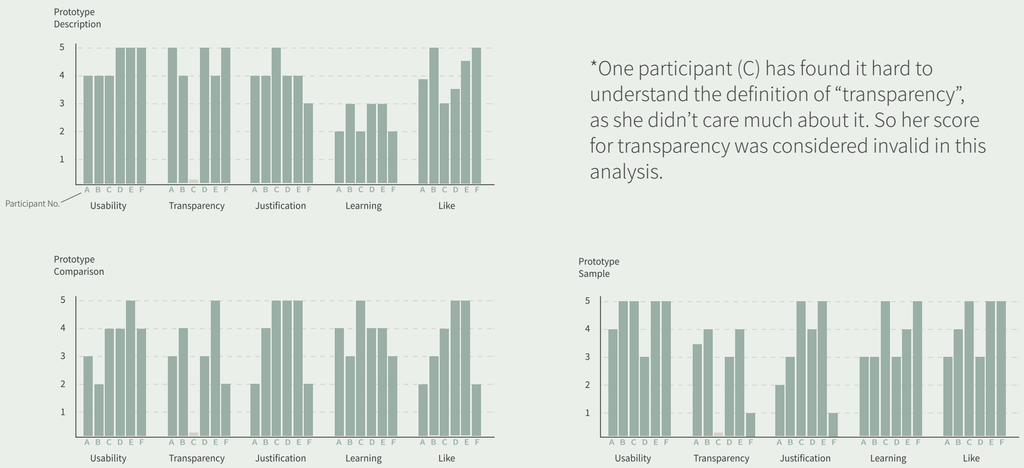


Figure 5-16. Radar chart of prototypes' average score



*One participant (C) has found it hard to understand the definition of "transparency", as she didn't care much about it. So her score for transparency was considered invalid in this analysis.

Figure 5-17. Overview of the rating result

By picking out those positive feedback relevant to the learning goal (shown in figure 5-18), we can see how different properties help them learn better.

To conclude, the different explanation properties have facilitated birders' learning in the following ways:

- Teach them what characterizes a certain species of birds. (Highlights on image and the textual description)
- Show them what is the contrast between two similar bird species. (Comparison)
- Deepen their impression of the bird's appearance. (Sample photos)

Properties	In which prototype	Quotes	Who held a similar opinion (No.)
Highlights on the prediction photos	"Description"	*"It taught me where to look at when seeing these birds."*	C D E
Comparison between different prediction results	"Comparison"	*"This feature is useful for learning to distinguish birds."*	A C D E F
Showing sample photos	"Sample"	*"Checking sample pictures is helpful in identifying and learning birds."*	A C E F
		"Showing a large number of sample pictures is a bit overwhelming for the beginners, but very helpful for advanced users."	A B C D E F

Figure 5-18. Positive feedback on how different properties facilitated learning

RQ9: What could motivate the target users to take part in the annotation process?

5.5.3 Attitude towards making annotations

At the end of this evaluation session, participants were asked how likely they would participate in the annotation tasks.

Their answers lead to two design strategies: achievement-driven and learning-driven. The former would take the form of making their contribution

Quotes	Who held a similar opinion (No.)
"I hope the queries will not be too open, because otherwise, I don't know what to fill in."	A
"I would love to do it with an appealing reward system."	D
"Doing this will provide me with a sense of achievement and contribute to the birder community and enable myself to learn at the same time."	D E F
"Be careful that not all the people are qualified to make the annotations, so you'd better test them first."	F

Figure 5-19. Participants' comments and advice on making annotation

apparent to them, so that participants would participate in the annotation process with the belief that they are helping the community. The latter proposes strategies to motivate participants' participation by allowing them to gain knowledge from the annotating activity.

One of the reasons why complete novice participants didn't want to perform it was their concern about the difficulty of the activities, which gave us insight into how to make the process more user-friendly for complete novices.

Accordingly, we can draw the following recommendations for developing an annotation process:

- Test the annotators in advance to make sure they are qualified.
- Find ways that are intriguing for birders and enable learning.
- The request should not be too open-ended, i.e. provide enough guidance especially for entrance-level annotators.

RQ8: What else do they expect from an educational bird app with explanations?

5.5.4 Ideas on providing explanation

We also gathered the participants' thoughts on how to create product features, which covered a wide range of topics (see apendix I). Those on information representation and backend technology were left out, and the remainder were summarized in figure 5-20.

The ideas of "showing more specific features" and "showing birds of various ages/genders/molting phases" among them illustrate what the participants want to learn from such a product.

Category	Ideas	Quotes	Proposed by
General	Show more detailed characteristics of the birds (like pattern, shape).	*"Some subtle differences were not described, for example, the nuance of the bill shape." -A*	A B C
Input method	Input features manually as a supplement to the input of photos.	*"I'd like to use the feature input function as a supplement to the image information. Some information is hard to recognize on the image." -B*	B D
	Modification of the influence of ambient light.	*"It will be great if the technology can modify the ambient light and angle of the photos." -D*	D E
Showing prediction	Present species of the same genus as reference.	*"... it would be of great help if you also link to the other 2 kinds of blue magpie here, which helps to determine a potential range." -E*	E
Showing explanation	Show sample pictures of the predicted species of different ages, gender, postures, molting stages, to present various appearances.	*"(I want to have) filters for location, ages, genders of the birds." -D*	C D E F

Figure 5-20. Participants' ideation on the product features

Though not all of these thoughts are relevant to the research scope, they will be interesting recommendations for other bird ID apps developers.

5.6 Discussion on limitations

In this evaluation, participants were tested with prototypes that are not functioning, and the incompleteness of prototypes' functions may influence the evaluation outcome in the following facets.

5.6.1 The ranking of explanation goals

The responders were confused by the ranking question during the evaluation. It may be argued that these three objectives interact with one another, making them difficult to rank in terms of "importance."

Transparency, for example, leads to justification, and both contribute to learning. Some may consider "transparency" to be the most important goal since it is the foundation, while others may consider "learning" to be the most important goal because it is the final goal.

However, by showing the prototypes and asking this question, we were able to learn how they saw these various goals, as well as the link between them. And, when utilizing an explainable bird ID tool, we could state that "learning" was the ultimate goal for the birders.

5.6.2 Impact of the "imperfect" target photo

As indicated in the participant demographic (5.4.3), not every participant knows the bird ahead of time (especially those from entrance level). As a result, the information on the target picture, such as the feather color altered by the sunshine, misleads them. However, participants stated that if they had seen the bird in person, they would have been able to tell the feature color more correctly, implying that their real-life experience with the app will differ from that of this evaluation session.

5.6.3 Limited information around the predicted species

In the primary prototypes, only limited information and interaction were shown, which influenced the experience in some ways. The participants would have to imagine the information shown by the envisioned web app, making it difficult for the users at the entrance level to justify the predictions made by the computer.

Summing up Chapter 5

In this chapter, we aligned the capabilities of the model explanation and the demands of the target users by evaluating primary explanation mock-ups with 6 birders. We have found that learning appeared to be the goal that most participants pursued while checking the provided explanations. Besides, we gain insights on how different properties in the explanations facilitate people in learning to identify birds.

In addition, the birders have contributed recommendations for improving the usability and some new ideas on the product's features, some of which will be carried on to the final prototypes.

Takeaways

- Learning was the most important, or rather, the ultimate goal of using a bird ID app with explanations for our target users.
- The explanation on bird species classification results can help the entrance bird hobbyists in knowing where to focus when recognizing birds, and in knowing the contrast between two similar birds.
- In terms of appearance, the users want to learn about the subtle differences that characterize different species, and the different appearances within species, which is currently missing in the explanation prototypes.
- The users would like to annotate out of learning purpose or sense of achievement.

CHAPTER 6.
TESTING THE ANNOTATION PROCESS

Main RQ: To what extent are the end-users able to make annotations correctly on the photos pre-processed?

In order to validate the underlying hypothesis of this project, a test with real annotation tasks will be conducted at this stage. Previously, prototypes were presented to the participants showing, to some extent, an ideal situation. During the annotation test of this chapter, the participants will be presented with real materials that come out directly from the bird species classification models.

By doing this, it will be found out to what extent the target users are able to make annotations correctly on those real materials.

6.1 Background

We looked at what the participants need to see about the explanation for their learning in the previous chapter by presenting them with explanation prototypes in a theoretical setting.

In this chapter, we'll look at how the approach works in terms of involving end-users in the annotation process at this point. Now we'll run the test in actual environments, using original testing materials derived directly from a real classification model.

Main Research Questions

RQ10: To what extent are the end-users able to make annotations correctly on the photos with bounding boxes?

a. To what extent are the end-users able to make annotation correctly on photos that the model found hard to classify?

With the setting of quantitative research, we can also validate the hypothesis that the annotation process itself can help people learn about birds, or learn about identifying bird species, to be specific. So here comes our minor research question at this stage:

Minor Research Questions

RQ11: Does making the annotation enable the end-users to be better / more confident at telling birds apart?

6.2 Method

6.2.1 The hypotheses

In response to the research questions, the following hypotheses were drawn.

Hypothesis 1a: *The participants can make annotations with high accuracy on the testing bird photos with bounding boxes.*

Hypothesis 1b: *The participants can make annotations with high accuracy even on the photos that are difficult to classify for the model.*

Hypothesis 2a**:** *The participants can do better in telling apart the birds after completing the annotation tasks.*

Hypothesis 2b: *The participants will be more confident in telling birds apart after completing the annotation tasks.*

The user test will be designed to collect data to validate the hypotheses.

6.2.2 Set-up of the user tests

This test will be carried out in the form of an online survey, as it is the easiest way I found to collect annotations from the participants. I chose the Qualtrics platform to run the survey.

To enable understanding of the general users, professional wordings were avoided, with words like “classification” and “annotation” separately replaced by “identification” and “description”.

The American Goldfinch and the Lesser Goldfinch were chosen as the two species to be distinguished in this test.

The overall structure includes:

(A= the American Goldfinch, B=the Lesser Goldfinch)

Introduction

- An introductory diagram showing the main differences between A&B
- Show sample photos (4A+4B)

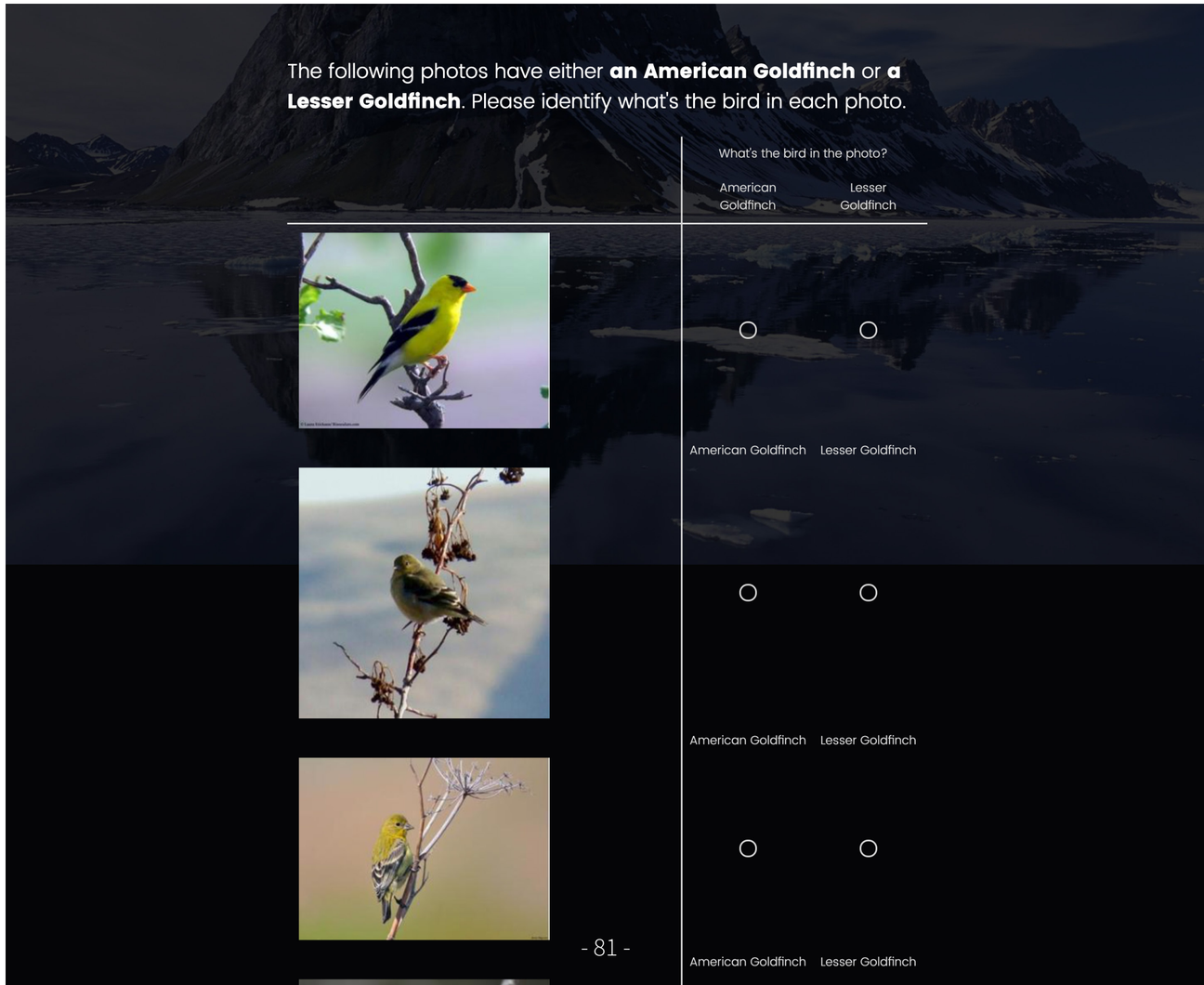
Identification task I (pre-test)

- Exercises: Identification exercise with 4 bird photos (2A+2B). Show the correct answer after submission.
- Identifying birds on 10 bird photos (5A+5B)
- Feedback question: *“How confident were you in the identifications you made just now?”(5-scale Likert)*

Descriptions task

- Instructions and examples
- Exercise*2: Description exercise on 2 photos(1A+1B) with bounding boxes. Show the correct answer after submission.
- Description task on a photo of A*2
- Description task on a photo of B*2
- Feedback questions:

Figure 6-1. Screenshot of the Identification Task screen



"How clear were the description tasks for you?"(5-scale Likert)

"How confident were you in the descriptions you made?"(5-scale Likert)

"How easy is it for you to identify the body parts of the birds?"(5-scale Likert)

"How easy is it for you to identify the color of the highlighted areas?"(5-scale Likert)

"What did you find hard about the description tasks?"(text entry)

Identification task II(post-test)

- Identifying birds on 10 bird photos (5A+5B)
- Feedback questions: *"How confident were you in the identifications you made just now?"(5-scale Likert)*

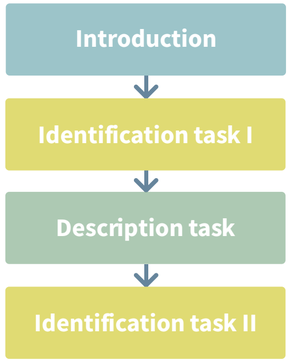
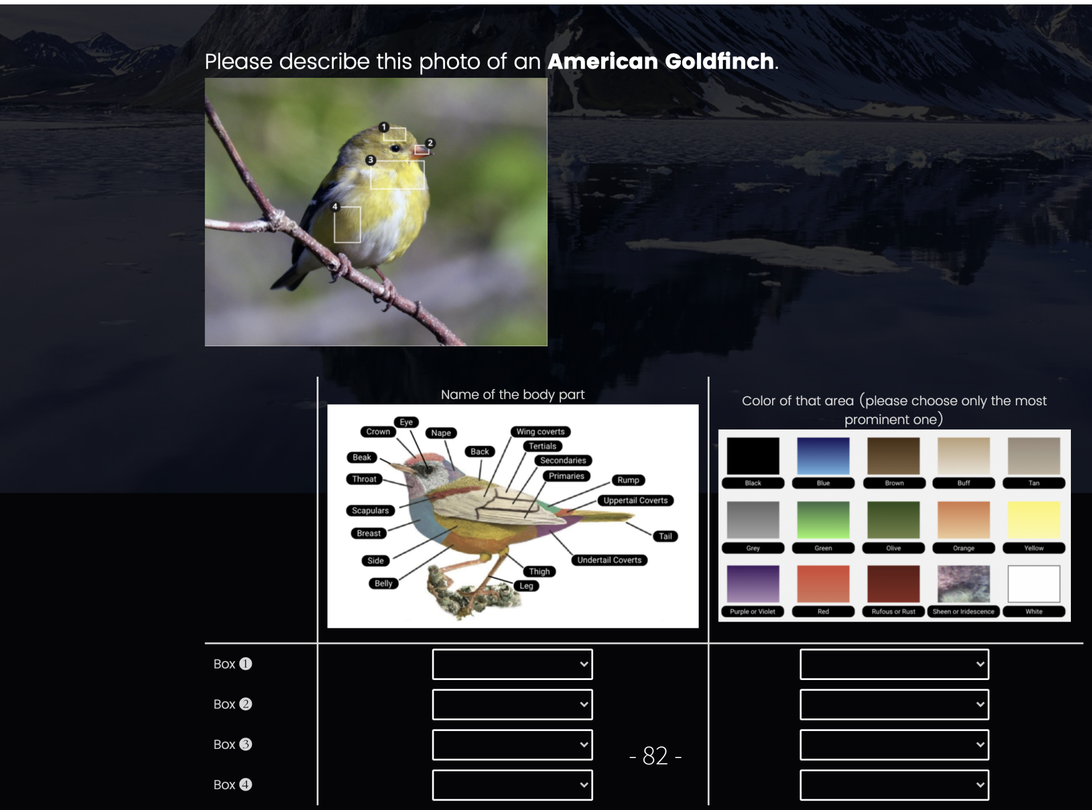


Figure 6-2. The test flow

After setting up the online test, one pilot test was conducted among a user, after which few adjustments were made to improve the usability of the formal test.

Figure 6-3. Screenshot of the Description Task screen



6.2.3 Testing materials

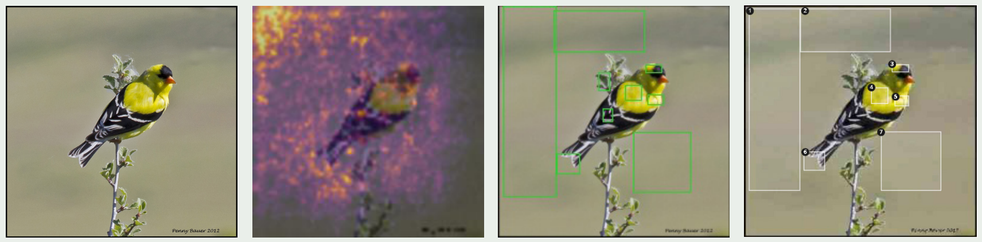
To simulate how the annotation process goes on in the wild, data from a dataset and a machine learning bird classification model were used to build this test, which was both provided by the SECA developers.

The dataset contains photos of 10 species of birds with both training and testing photos(1,291 photos for train, 1,470 photos for test). Besides, the classification model that could identify 10 birds was trained using the mentioned dataset.

With the data photos and the classification model, saliency maps were generated highlighting the important areas for the model to make its decision.

Originally, the bounding boxes couldn't be automatically generated and required humans to draw them, too. But to simplify the process a bit at this stage, bounding boxes were pre-drawn by the SECA developers on the testing photos based on the saliency maps, for human annotators to later annotate on (figure 6-4 (c)).

The confidence scores by the model (figure 6-5) were used to decide which photos in the dataset will be used for tests. For example, we want the confidence values of photographs for Identification tasks I and II to be almost identical, such that the complexity of these two tasks does not affect people's accuracy.



a) Original photo in training dataset | b) The test photo with saliency map | c) Photo with bounding boxes drawn by the developers based on the saliency map | d) Photo with simplified bounding boxes for tests

Figure 6-4. Proccessing test materials from original dataset

image_name	category	predicted	confidence
fe63ca47afa74222a0a5d9639ee44e8.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9998030066490173, 'bufflehead': 1.1346950259394362e-06, 'downy_woodpecker': 2.3198184706599e-08, 'gila_woodpecker': 1.2872140020547818e-09, 'hairy_woodpecker': 4.197574926534865e-11, 'hooded_merganser': 6.400919239091252e-10, 'lesser_goldfinch': 0.00019490173144737, 'mandarin_duck': 4.284264942810978e-09, 'mink_parakeet': 4.215471527402315e-09, 'pine_grosbeak': 9.37946721171809e-07}
1ead7bb8b2a54e9e9952e6d6bb9d9e81.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9982343912124634, 'bufflehead': 4.199493730538961e-08, 'downy_woodpecker': 4.242521754349582e-06, 'gila_woodpecker': 8.622811037639622e-06, 'hairy_woodpecker': 4.013114583662385e-07, 'hooded_merganser': 7.191866302491699e-08, 'lesser_goldfinch': 0.0007492591383457 18, 'mandarin_duck': 5.135770493025253e-07, 'mink_parakeet': 0.0001008494407869875 4, 'pine_grosbeak': 1.1964880641812674e-08}
8962d575e304eaac8479110e6b899cae.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.6683021187782288, 'bufflehead': 0.0002600362604080554 25, 'downy_woodpecker': 0.03339593484997749, 'gila_woodpecker': 0.0001962131671793 76, 'hairy_woodpecker': 0.0001920966350599927, 'hooded_merganser': 0.0007127708529292916, 'lesser_goldfinch': 0.2256999648310702 5, 'mandarin_duck': 1.4450646531202424e-05, 'mink_parakeet': 0.0000641885912418 4, 'pine_grosbeak': 0.0015949618771332502}
3fb41eddd5da4066aa1373dd844214be.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9999842563737793, 'bufflehead': 1.5220477202941409e-08, 'downy_woodpecker': 1.9951202832055262e-09, 'gila_woodpecker': 6.8869864322187235e-09, 'hairy_woodpecker': 4.786682553548892e-07, 'hooded_merganser': 7.549863312306115e-08, 'lesser_goldfinch': 1.4978756553318817e-05, 'mandarin_duck': 1.1009547762341754e-07, 'mink_parakeet': 1.4921666124934063e-09, 'pine_grosbeak': 3.2031392055202104e-08}
1fe0cf8e94644db0a3d098702ae298.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9971908820963806, 'bufflehead': 3.701864272898092e-07, 'downy_woodpecker': 3.7525624065897277e-05, 'gila_woodpecker': 0.0001473815929892695, 'hairy_woodpecker': 2.5881705369101837e-05, 'hooded_merganser': 1.3262418008253007e-07, 'lesser_goldfinch': 0.0025673853233455661, 'mandarin_duck': 2.2744211776295933e-08, 'mink_parakeet': 1.0516118891246151e-05, 'pine_grosbeak': 1.9533181116681096e-05}
323f091fc4694664930735bbd7e2edcb.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9995393753051758, 'bufflehead': 6.421269915790617e-08, 'downy_woodpecker': 1.300779331359081e-05, 'gila_woodpecker': 6.81747478665784e-05, 'hairy_woodpecker': 3.906390588781505 4e-07, 'hooded_merganser': 1.1701208002534713e-08, 'lesser_goldfinch': 0.0003763940476346761, 'mandarin_duck': 7.337098395510111e-08, 'mink_parakeet': 2.6133500341529725e-06, 'pine_grosbeak': 3.988664937894504 4e-08}
c162fe6f60f94cb7908979e9b6993d2d.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.39286994934082 03, 'bufflehead': 0.037307862192392 35, 'downy_woodpecker': 0.026694567874072982, 'gila_woodpecker': 0.009039835073053837, 'hairy_woodpecker': 0.018356033834566993, 'hooded_merganser': 0.014174719703096061, 'lesser_goldfinch': 0.31508679249902466, 'mandarin_duck': 0.0045607641758276 39, 'mink_parakeet': 0.17313766479492188, 'pine_grosbeak': 0.005705383140593767}
3c9d0a84fb194c8dbdde2385e71afa7d.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9995797276496887, 'bufflehead': 3.772157924686325 7e-06, 'downy_woodpecker': 1.0735153409768827e-05, 'gila_woodpecker': 5.468847227247554e-07, 'hairy_woodpecker': 2.0149780084466 33e-06, 'hooded_merganser': 1.6076070323833849e-06, 'lesser_goldfinch': 0.0002002778309381104, 'mandarin_duck': 6.397232071509879e-06, 'mink_parakeet': 3.9065196286300 2e-05, 'pine_grosbeak': 0.0001560042001074 18}
a8017b6f6c1948c0853cde03aacc8e92.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.7795418092000269, 'bufflehead': 5.109861376995335e-06, 'downy_woodpecker': 0.0042501183065980673, 'gila_woodpecker': 0.0009636257831007 24, 'hairy_woodpecker': 3.8522584873132e-06, 'hooded_merganser': 4.665769665734842e-06, 'lesser_goldfinch': 0.21531538665294647, 'mandarin_duck': 1.8408880464448885e-07, 'mink_parakeet': 1.9032297871035553e-07, 'pine_grosbeak': 8.229029037127271e-05}
8eca02ad10c145448dc13e561fdc5d83.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9888365722932379, 'bufflehead': 1.1432608516770415e-05, 'downy_woodpecker': 0.0005840731319904006, 'gila_woodpecker': 7.182943591033109e-07, 'hairy_woodpecker': 0.0004714081006588544, 'hooded_merganser': 2.7147640139446594e-05, 'lesser_goldfinch': 0.0098426425852952976, 'mandarin_duck': 1.6855372450663708e-05, 'mink_parakeet': 0.0002214951673522591 6, 'pine_grosbeak': 1.5674464520988e-05}
ffb950c50604fe5b845c5e620f98bd2d.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9999901056289673, 'bufflehead': 1.8435081239331907e-09, 'downy_woodpecker': 2.4978037416190 09e-09, 'gila_woodpecker': 1.2821287498354654e-07, 'hairy_woodpecker': 2.7454191697806607e-10, 'hooded_merganser': 9.336072359644731e-09, 'lesser_goldfinch': 3.387354809092357 8e-06, 'mandarin_duck': 1.1090470763974736e-09, 'mink_parakeet': 6.432939244405134e-06, 'pine_grosbeak': 1.1267328921586 02e-06}
375769d5dc7440fba8d8ff829c4370a8.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.5113171339035034, 'bufflehead': 2.5065002517976614e-08, 'downy_woodpecker': 7.093709552370356e-07, 'gila_woodpecker': 6.037074484765981e-08, 'hairy_woodpecker': 1.3440818204185746e-05, 'hooded_merganser': 0.0006922538741491735, 'lesser_goldfinch': 0.4879713356494936, 'mandarin_duck': 4.807052916976503e-06, 'mink_parakeet': 2.0551327027035617e-08, 'pine_grosbeak': 2.1267514682676847e-07}
b6e676dbe39e46988bd33358bffa302e.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9967522621154785, 'bufflehead': 7.550888742519019e-10, 'downy_woodpecker': 1.3400204807112439e-05, 'gila_woodpecker': 1.4290037562197617e-09, 'hairy_woodpecker': 7.737509166449748e-06, 'hooded_merganser': 1.4074318241838114e-07, 'lesser_goldfinch': 0.0031357142429514 8, 'mandarin_duck': 1.9884666307939854e-07, 'mink_parakeet': 0.029121429193795e-05, 'pine_grosbeak': 0.0496991159873853e-07}
6cfcf1541cbb4f408129bd24ab538d3.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.9998805518577576, 'bufflehead': 6.017539533331728e-08, 'downy_woodpecker': 3.783274848956353e-08, 'gila_woodpecker': 1.6582473241766316e-11, 'hairy_woodpecker': 1.1782987088039753e-10, 'hooded_merganser': 6.380258987837 0e-09, 'lesser_goldfinch': 1.5641180652892217e-05, 'mandarin_duck': 7.144819136328806e-08, 'mink_parakeet': 0.0001474192687175754 4, 'pine_grosbeak': 1.1443795528975897e-06}
b2b741572ff84dbd6bd882acad14f7ba4.jpg	american_goldfinch	american_goldfinch	{'american_goldfinch': 0.8581004142761 23, 'bufflehead': 0.000757173052988464, 'downy_woodpecker': 0.0095925228670233 45, 'gila_woodpecker': 0.0026941932737823, 'hairy_woodpecker':

Figure 6-5. (Incomplete) table of confidence scores of predictions for every training photos

To conclude the materials coming from the model include:

1) Original photos in the test dataset;

2) Photos with saliency maps;

3) Photos with bounding boxes made by the developers;

4) Annotations made by the developers (as baseline);

5) Confidence scores of all the predictions by the classification model.

6.2.4 Participants selection

According to the positioning of the target users, there were no specific requirements for the participants of this test. We aimed at collecting at least 8 responses for the result to be convincing.

We posted the link to this survey with a brief introduction on /SurveyExchange on Reddit and Douban (Chinese version of Reddit), and finally got 16 valid responses back (1 from the pilot test was excluded).

Besides, we invited a participant to complete the whole test in front of me, and asked some questions after, to observe what were the problems that she encountered.

6.3 Procedure

6.3.1 Data collection and analysis

Both quantitative and qualitative data were collected from the online survey for the analysis.

1. Correctness of the description (annotation) tasks

Data to collect:

- Calculate the correctness of the annotation made by the participants, using the annotation made by the researcher and the developer as baseline (leave room for the grey area).

To evaluate the quality of the collected annotations (***hypothesis 1a and 1b***), the correctness was calculated using the annotations made by the researcher and the developer as the baseline.

In particular, we input what we believe are the proper answers as well as certain "acceptable alternatives" in the background setting of the questionnaire platform. For example, "throat" and "belly" are acceptable answers for the target bird's breast area. The rationale for this is that the location of the bounding boxes can be ambiguous at times, and collecting these acceptable answers could aid in the creation of explainability as well as make learning easier for end-users.

Besides, the annotations on colors were left out in this correctness analysis for two reasons: firstly, it was hard to decide which was the “correct” color name as they were sensed differently by different people and could be influenced by ambient light or shadow; secondly, previous research shows (Chapter 4&5) names of color played a less important role in people’s learning process, compared to that of body parts.

Then, out of the total number of inquiries, we calculate how many matches there are in the participants' responses to determine the annotation accuracy.

2.Feedback on the description task

Data to collect:

- Clearness of the tasks for the participants (measured by 5-point Likert)

- The difficulty of the tasks for the participants (measured by 5-point Likert)

- Reasons why they found it unclear/ difficult

This part of the analysis was about how difficult the participants found it about making the annotation based on their own feedback.

3. Comparison of the participants’ confidence in completing two identification tasks

Data to collect:

- Confidence the participants have for each identification task (measured by 5-point Likert).

Participants were asked directly how confident they were after completing two identification tasks. This was to provide qualitative data to see whether the annotation process makes them more confident in distinguishing the birds (**hypothesis 2a**).

4. Comparison of the correctness of two identification(classification) tasks (before and after)

Data to collect:

- The average correctness data of identification task I and identification task II.

When the average correctness of identification task I is compared to the average correctness of identification task II, quantitative evidence will be presented on whether the annotation process aids participants in distinguishing birds.

If correctness II is significantly higher than correctness I, then ***hypothesis 2b*** is proved, which means the participants can do better in telling apart the birds after completing the annotation tasks.

6.4 Findings

6.4.1 Correctness of the annotation

Main Research Questions

RQ10: To what extent are the end-users able to make annotations correctly on the photos with bounding boxes?

a. To what extent are the end-users able to make annotation correctly on photos that the model found hard to classify?

The graphic in figure 6-6 shows their correctness for each description task.

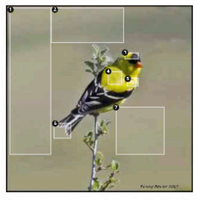
The second and fourth photo were the ones with lower classification confidence, meaning that the identification model found them hard to classify. The test result shows that those photos with lower confidence were a bit more difficult for humans to annotate, too, with the accuracy of 92% and 93%, compared to the accuracy of 95% and 94% on the easier ones.

Generally speaking, participants were able to make the annotations with high correctness (correctness more than 92%). ***Hypothesis 1a*** and ***1b*** were verified.

Collected annotations on body parts

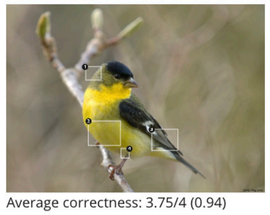
Box❶ Crown(16)
Box❷ Beak(16)
Box❸ Throat(8), Breast(6), Scapulars(2)
Box❹ Side(13), Belly(2), Breast(1)

Average correctness: 3.81/4 (0.95)



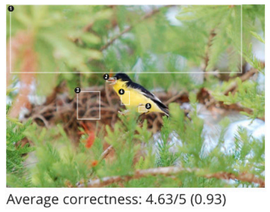
Box❶ Non-bird area(16)
Box❷ Non-bird area(16)
Box❸ Crown(16)
Box❹ Back(10), Scapulars(3), Side(2), Breast(1)
Box❺ Breast(7), Scapulars(6), Throat(1), Belly(1), Wing coverts(1)
Box❻ Tail(16)
Box❼ Non-bird area(16)

Average correctness: 6.44/7 (0.92)



Box❶ Nape(16)
Box❷ Belly(13), Breast(3)
Box❸ Primaries(15), Secondaries(1)
Box❹ Thigh(16)

Average correctness: 3.75/4 (0.94)



Box❶ Non-bird area(16)
Box❷ Beak(15), Throat(1)
Box❸ Non-bird area(16)
Box❹ Breast(16), Side(1)
Box❺ Belly(12), Undertail coverts(1), Thigh(3)

Average correctness: 4.63/5 (0.93)

Figure 6-6. Annotations collected from the participants and their correctness

Moreover, it is found from the test result that the most popular description collected from the participant is 100% correct. This finding indicates that if we pick out the most frequently made annotations from the collection, we can be sure to get the right ones.

6.4.2 Feedback on the annotation process

Overall, the description task was clear to them and participants felt relatively confident about the description they made. It was revealed by the result that 87.5% of the participants found the description tasks somewhat clear or extremely clear (figure 6-7). And 75% of them felt moderately confident or very confident in the annotations they made (figure 6-8).

In general, we can say that the participants found the annotation task was not too difficult to complete.

Specifically, more participants found annotating the colors difficult compared to annotating the body parts, and this was mostly caused by the unclear instructions (figure 6-9, 6-10).

In the text entry question “what did you find difficult about the description task” (figure 6-11), some of the replies wrote:

Answer	%	Count
Extremely unclear	0.00%	0
Somewhat unclear	6.25%	1
Neither clear nor unclear	6.25%	1
Somewhat clear	50.00%	8
Extremely clear	37.50%	6
Total	100%	16

Figure 6-7. Result for *"Q29-How clear were the description tasks for you?"*

Answer	%	Count
Not confident at all	0.00%	0
Slightly confident	25.00%	4
Moderately confident	50.00%	8
Very confident	25.00%	4
Extremely confident	0.00%	0
Total	100%	16

Figure 6-8. Result for *"Q30-How confident were you in the description you made?"*

"Especially for the last lesser goldfinch, the bird in the image was far away and its parts were hard to distinguish. I also had a bit of a hard time deciding between olive/tan/grey a few times."

"Bird in the diagram is not always the same shape/orientation as the bird in the photo."

"Choosing which two colors is dominant."

"Ambiguity about which area was meant"

To conclude, the reasons that made the description tasks difficult for them were (concluded from their feedback, see in figure 6-11):

(Annotating colors)

1) Choosing one dominant color out of multiple;

2) Deciding between similar colors;

3) Photos were too small;

(Annotating body parts)

4) Photos were too small;

5) More than one body part were included in the highlighted area;

6) Birds in the photos were not in the same shape/orientation as that were in the reference diagram.

Answer	%	Count
Extremely easy	6.25%	1
Somewhat easy	68.75%	11
Neither easy nor difficult	6.25%	1
Somewhat difficult	18.75%	3
Extremely difficult	0.00%	0
Total	100%	16

Figure 6-9. Result for *"Q27-How easy is it for you to identify the body parts of the birds?"*

Answer	%	Count
Extremely easy	0.00%	0
Somewhat easy	56.25%	9
Neither easy nor difficult	12.50%	2
Somewhat difficult	31.25%	5
Extremely difficult	0.00%	0
Total	100%	16

Figure 6-10 Result for *"Q28-How easy is it for you to identify the colors of the highlighted areas?"*

What did you find difficult about the description task?
finding the part in the list and picking a colour when there were to in equal proportions were the most difficult things, but still very easy
Hard to tell the body parts from one to another, especially for the tiny ones; and difficulties in color categorization as no lateral comparison showed as visual hints.
Choosing which two colors is dominant
bird in the diagram is not always the same shape/orientation as the bird in the photo
especially for the last lesser goldfinch, the bird on the image was far away and its parts were hard to distinguish. i also had of bit of a hard time deciding between olive/tan/grey a few times
ambiguity about which area was meant
some colors felt blended or unclear because the bird was small
some parts of the nody that we have to discribe have more than one color, but I can't choose both.
Identify colors
Some of the colors are very similar, like tan and buff or olive and tan and it's hard to differentiate.

Figure 6-11 Result for *"Q45-What did you find most difficult about the description task?"*

6.4.3 Comparison of the first and second identification correctness and confidence

Minor Research Questions

RQ11: Does making the annotation enable the end-users to be better / more confident at telling birds apart?

When comparing the participants' confidence level for two identification tasks, some of their confidence dropped, some rose. When converted into numeric values (figure 6-12), their average confidence level even dropped a little bit (from 2.44 to 2.19, 1=not confident at all, 5=extremely confident). Conducting a T-test on the data, got p=0.388286, which means the decrease in confidence was not significant (***hypothesis 2b*** was not verified) (the result is significant when p<0.05).

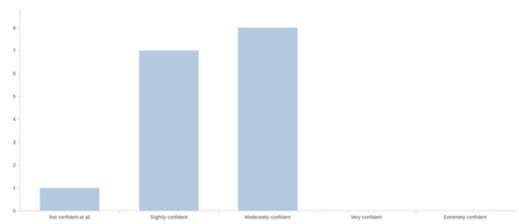
A rise was noticed in the average correctness of identification tasks from 6.63 to 7.56 (out of 10) (figure 6-13). However, through a T-test, no significance was found in the comparison of these two scores (p=0.152696). In conclusion, the current data couldn't lead to a conclusion on whether the annotation tasks could help people in identifying birds from this test (***hypothesis 2a*** not verified).

Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
Identification1	3.00	9.00	6.63	1.65	2.73	16

Field	Minimum	Maximum	Mean	Std Deviation	Variance	Count
Identification2	4.00	10.00	7.56	2.03	4.12	16

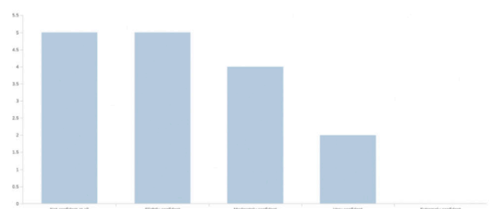
Figure 6-12. Correctness of identification task I (above) and identification task II (bottom)

Participants' confidence in identification task I



Answer	%	Count
Not confident at all	6.25%	1
Slightly confident	43.75%	7
Moderately confident	50.00%	8
Very confident	0.00%	0
Extremely confident	0.00%	0
Total	100%	16

Participants' confidence in identification task II



Answer	%	Count
Not confident at all	31.25%	5
Slightly confident	31.25%	5
Moderately confident	25.00%	4
Very confident	12.50%	2
Extremely confident	0.00%	0
Total	100%	16

Figure 6-13. Participants confidence in identification task I (left) and identification task II (right)

6.5 Discussions

In this test, we used the materials that came directly from a real bird classification model to simulate a real annotating process. Instead of letting the participants write the description themselves, we provide them with reference diagrams from which they can choose the descriptive words.

The test result verified our first hypothesis that the participants were able to make annotations with high correctness (more than 93% on average) this way, and we can be sure to get the right annotation if we pick out the most frequent ones.

However, it remains unsure whether they can do better in telling birds apart as the test result shows no significance in the comparison of two identification tasks, which may be caused by the the following factors:

1) The annotation process wasn't deliberately designed for teaching, but for collecting annotation instead, which limited the likelihood for participants to learn from it;

2) The chosen platform we employed for the test had (qualtrics) quite some limitations in terms of the interface design, which potentially influenced the participants' cognition of the presented information;

3) The sample size of the test wasn't large enough to notice any significant differences in participants' confidence level and accuracy for the two classification tasks.

Summing up Chapter 6

This chapter conducted a scientific study on whether we can collect annotations needed by the explainable machine learning model from the target users. Along with this goal, we also wanted to find out whether the target users can be better at identifying birds after completing the annotation tasks.

The first goal was reached as we found out it is possible to get correct annotations out of what they made. And we also gained feedback on what they found hard about the annotation tasks, for usability improvements.

But for the second goal, no conclusion can be reached yet as to whether the annotation helps them in distinguishing birds out of the data we got from this stage.

Takeaways

- It was found out that participants were able to make annotations with correctness of more than 93%.
- It was found out in the test that the mostly picked descriptions were always correct, which could inspire the developer when collecting annotations.
- Participants found annotating colors were more difficult compared to body parts. The difficulty in annotating body parts was mainly caused by unusual angles or small figures.
- Participants' confidence and correctness in identification didn't improve significantly after annotation tasks.

CHAPTER 7.
TESTING THE ANNOTATION INTERFACES

Main RQ: How to design an interface that people would like to use for learning and making annotations?

In the SECA framework, annotations need to be collected from humans to enable semantic explanations of the visual features. Thus, this chapter aims at involving end-users in the annotation process, in a way that is engaging and also educational for them.

A prototype was developed using the materials coming out from a real classification model, in order to simulate a SECA process in the wild. Then, in the user tests, bird hobbyists will be invited to complete some tasks under the guidance on the prototype interfaces. In this way, the researcher seeks to find out most proper ways to engage users in the annotation process with game-like interactions and interfaces.

Research Questions

RQ12: Do the users find it engaging to do the annotation tasks through the game-like interaction and interfaces?

RQ13: Do the users enjoy learning about birds through the prototyped interaction and interfaces?

RQ14: How is the usability of the interfaces and what can be improved?

7.1 Design considerations

7.1.1 Raw materials

In this interface test, we used the same 10-species identification model as was used in Chapter 6.

The materials coming from the model include:

1) Original photos in the test dataset;

2) Photos with saliency maps;

3) Photos with bounding boxes made by the developers;

4) Annotations made by the developers (as baseline);

5) Confusion matrix of the model.

The confusion matrix of the classifcation model (figure 7-1) was used to decide which bird species would be used for comparison. We gained the information from the confusion matrix that the Pine Grosbeak and the Lesser Goldfinch are the two species that the model had found hard to distinguish from the American Goldfinch, so we included these two birds in the classification challenge within the tutorial on identifying American Goldfinch.

Different from Chapter 6 where we kept materials as original as possible, in this stage I made changes to the materials to be more user-friendly.

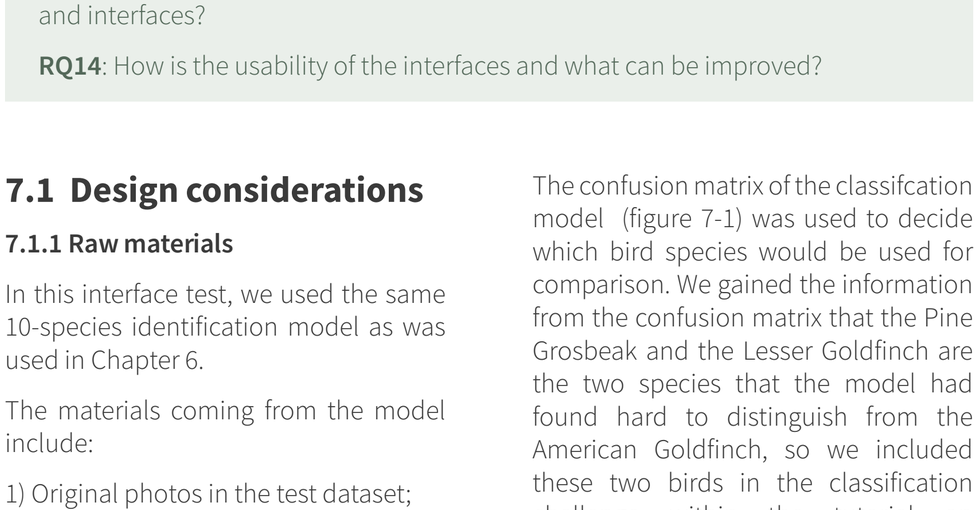
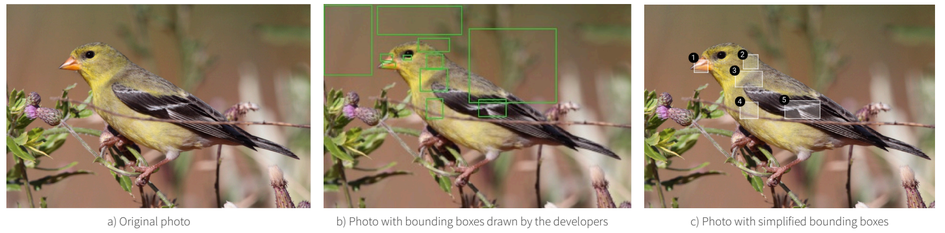


Figure 7-1. Confusion matrix of the classification model

Such editing includes:

1) Removing some bounding boxes from the photos on the introduction page;

2) Keeping only one box for each annotation task.



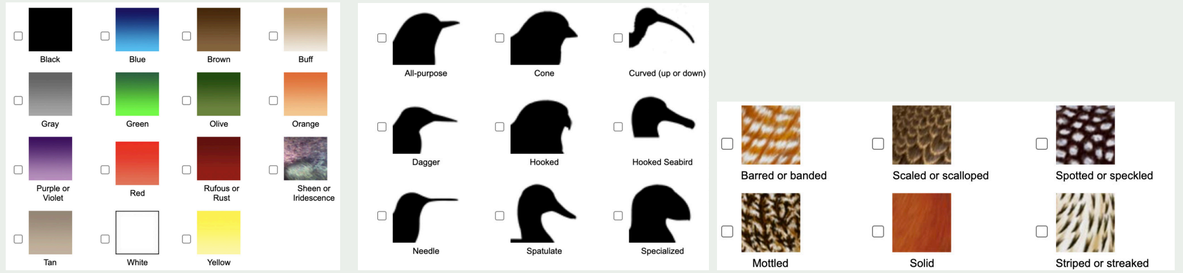
a) Original photo b) Photo with bounding boxes drawn by the developers c) Photo with simplified bounding boxes

Figure 7-2. Editing of test photos

7.1.2 Attributes

To guide the participants through the annotation process with consistent vocabulary, a list of words of body parts, colors, and patterns, etc, was provided as reference.

WhatBird.com is a website that guides people through bird identification which is briefly introduced in Chapter 1(1.1). Inspired by the image labeling process used by Branson, et al (Branson et al., 2010), we used the attributes provided by whatbird.com as options for the participants to select from, the attributes used for the prototypes included the name of colors, bill shapes and wing patterns.



a) Colors b) Bill shapes c) Body(back/belly/breast) pattern

Figure 7-3. Attributes provided by Whatbird.com

7.1.3 The bird topography diagram

A diagram of bird topography was presented to the users, reminding them of the name of each body part.

While there are different bird topography diagrams going around, the one made by Wild Bird Unlimited (figure 7-4) was chosen because:

1) The colored markings on the bird body will be easier for beginners to identify, compared to those solely with lines.

2) The terms it uses are not too difficult to be understood by beginners.

3) Meanwhile, enough details on nuance appearance are contained for when users get advanced.

4) The granularity of its description is proper, neither too fine-grained nor too coarse.

Aside from this one that was used as the main reference diagram, there were other diagrams that would be used as supplements to cover the field marks that are not included yet, depending on different contexts. For example, figure 7-5 showing markings on wings, would be used when the markings helped distinguish the target birds.

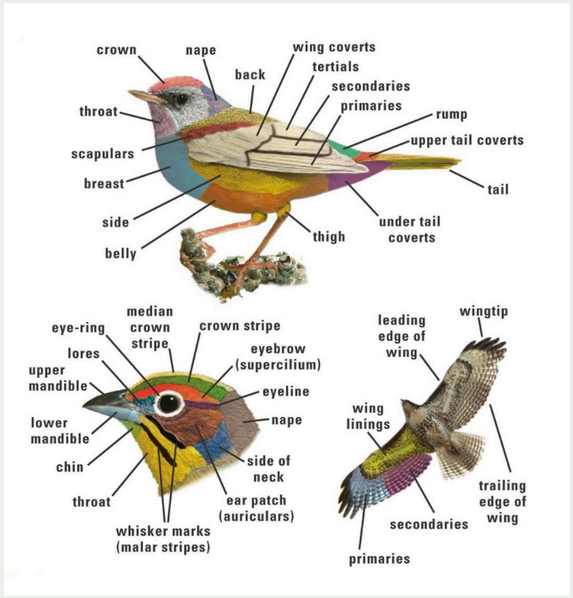


Figure 7-4. The bird topography diagrams

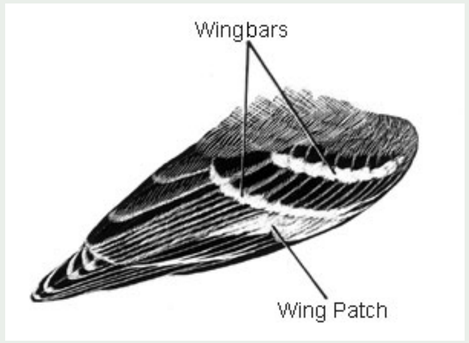


Figure 7-5. The markings on birds' wings (by John Schmitt/Cornell Lab)

7.1.4 Interfaces design

When designing the interface of the introduction and classification task page, we took into account the insights from explanation prototypes tests.

Interface of the characteristics introduction

At the beginning of the learning, we presented a screen that resembles the interface of the “Description” prototype in Chapter 5. This is to leave an impression on the users of characteristics of the target birds. (figure7-6)

Interface for the classification task

Comparison was considered helpful to learning by users in Chapter 5. So we kept the comparison part here, making it into a small test in between the annotation tasks. After users submitted their choice, the answer was shown with bounding boxes highlighting the contrast between the two birds. (figure7-7)

Interface for the annotation task

We wanted the users to choose from the offered vocabulary because of the lack of bird knowledge among the general public and the need for uniformity in the granularity of the expected annotations. In addition, each page just has one question for users to answer, ensuring a clear and simple experience. Users will be presented with an interface similar to figure 7-8 for annotation.

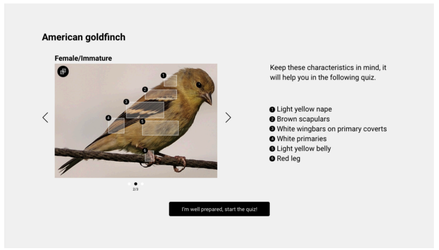


Figure 7-6. Characteristics introduction interface

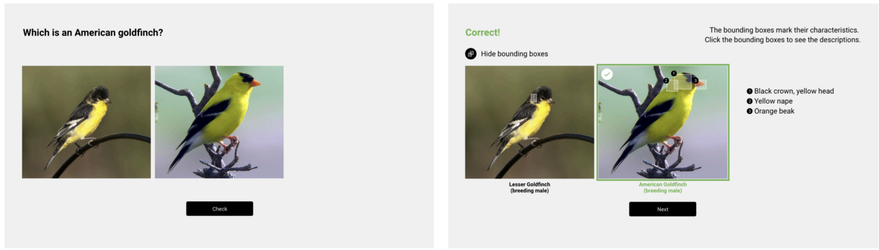


Figure 7-7. Classification task interface

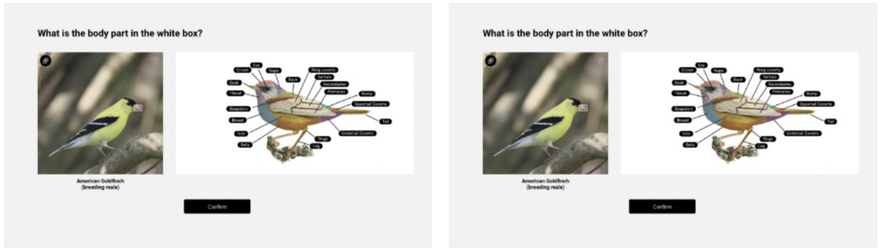


Figure 7-8. Annotation task interface

7.1.5 Overview of the test prototype

The test prototype consists of guidelines of annotations, introduction to the target birds, with few classification tasks and annotation tasks appearing alternatively.

Here's an overview of the prototype's setup.

1. Guidelines with examples

2. Introduction of characteristics

3. Classification task

- Distinguishing 2~3 bird

4. Annotation tasks:

- Annotating body part
- Annotating color
- Annotating shape of that part (bill, tail, wing)
- Annotating pattern of that part (Back, belly, breast, head, throat)

Check Appendix L to see the live preview and the page-by-page content of the test prototype.

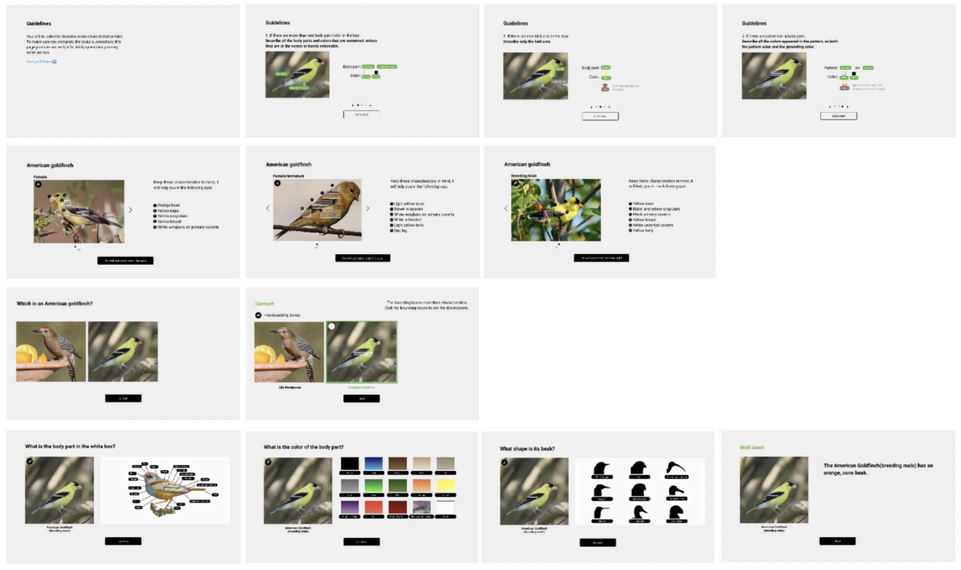


Figure 7-9. Screenshots of part of the prototype

7.2 Method

7.2.1 The evaluation questionnaire

After trying out the prototype, an online questionnaire will be filled by the participants.

The questionnaire includes a system usability scale (SUS)(Brooke, 1996) rating the ease of use, followed by questions evaluating separately the information representation, engagingness, and learning effect of the prototypes.(see Appendix M for the complete questionnaire).

At the end, the participants were asked to pick out the word and emoticon from the premo tool. (Desmet, 2018).

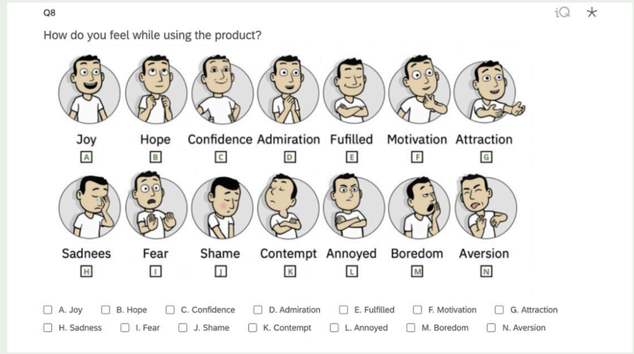


Figure 7-10. Screenshot of the questionnaire

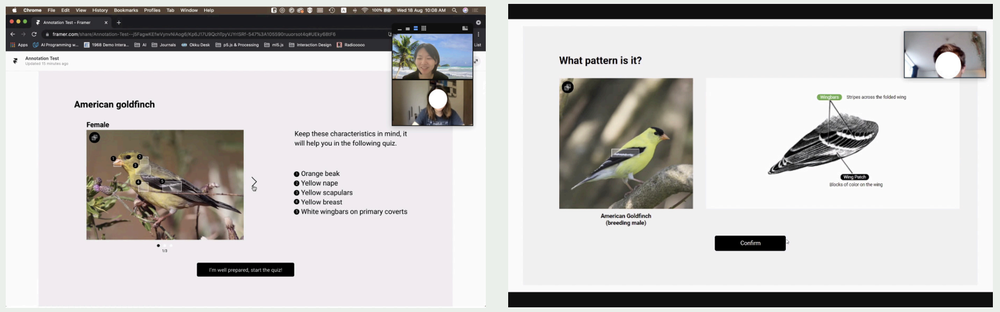


Figure 7-11. Captures of the test sessions

7.2.2 Participants selection

In Chapter 4, we decided the target users of this project to be people without much knowledge about birds. So in this test, I looked for participants who have either no previous bird identification knowledge at all, or only a little bit.

The participants (n=3) were recruited with snowball sampling from those who have shown interest in learning about birds. Two of them have no experience in bird identification, one has only a few months of birding experience. They will be referred to as P1, P2, P3 in the following text.

7.3 Procedure

7.3.1 The test protocal

The tests were conducted follow the following protocol:

1. Opening: introduction by the researcher to the process and content on informed consent (5 mins);

2. Participants freely tried out the prototype: (10 mins);

   a. Send participants the test link;

   b. Invite them to share their screen, and thinking out loud along the process;

3. Fill out evaluation questions (15 mins).

The tests were conducted via zoom and

were screen recorded with the permit of the participants.

7.3.2 Analysis

The Analysis was conducted based on the observation of the tasks completion and the results collected by the questionnaire. The complete result was documented in Appendix N.

Observation

During the tests, the researcher observed and noted down the participants' completion of each task with the description of "perfect success", "so-so success" and "fail".

And the problems they encountered were clustered into three typologies: "ambiguous guidance", "bugs", "unnoticed indication".

In this way, all the usability problems at different screens were documented for improvements.

P1, P2, P3: Participant number

- **Perfect success**
- **So-so success** (made nonfatal mistakes, or felt confused)
- **Fail** (didn't complete the task as expected)

a. Ambiguous guidance
b. Bugs
c. Unnoticed indication

	Completion of tasks	Types of problems	Annotation
Select	P1 P2 P3		
Guidelines(3)	P1 P2 P3	**a.**	
Intro(3)	P1 P2 P3	**a. c.**	
Classification-1	P1 P2 P3	**c.**	
Annotation-1	P1 P2 P3	**c.**	**P2: chose "short dagger" instead of "cone"**
Annotation-2	P1 P2 P3	**a.**	**P1 P2 P3: choose one color or both?**
Annotation-3	P1 P2 P3		
Classification-2	P1 P2 P3		
Classification-3	P1 P2 P3	**b.**	
Annotation-4	P1 P2 P3		**P3: not sure what to choose "breast"/ "throat"**
Annotation-5	P1 P2 P3		**P3: not sure what to choose "back"/ "nape"**

Figure 7-12.Task completion diagram based on observation

Questionnaire result

Besides the observation, the result collected from the questionnaire was used to analyze the participants' overall impression of the prototypes.

Metrics involved are the SUS score measuring the usability, how they rated the engagingness of the concept, and how they felt during trying out the prototype.

7.4 Findings

7.4.1 Engagingness of the prototype

RQ12: Do the users find it engaging to do the annotation tasks through the game-like interaction and interfaces?

In the evaluation, 2 out of 3 participants agreed that this kind of learning activity was engaging to them, and they rated 4 (out of 5) for the engagingness of the prototype.

While the other participant with a little bit more experience in birding didn't find it engaging, rating 1 (out of 5) for the engagingness.

> *"I would prefer a lighter way of learning. Now there are too many details in it which make me fear."-P3*

7.4.2 Overall feelings towards the experience

RQ13: Do the users enjoy learning about birds through the prototyped interaction and interfaces?

The first participant (P1) picked all the positive words (joy, hope, confidence, admiration, fulfilled, motivation, attraction) on the list to describe her feelings.

"The whole process is joyful. And I feel hope because it is a totally new thing to me. I feel admiration because I expected it to be very difficult to start but it wasn't. And fulfilled when the app told me that I was correct or accomplished something."-P1

P2 thought the experience has made him feel fulfilled, motivated and attracted, due to the representation of the interface and the game-like setting.

> *"I feel fulfilled because I learnt something. And the whole setting and that its content is getting deeper made me feel motivated and attracted."-P2*

P3 felt joy, confidence and fear. She thought the form of a game-like web-app seemed fun to her, and the tasks were easy. But still she feared making mistakes during the process.

> *"The learning was joyful and relaxing, but I felt insecure and feared that anything would go wrong during the usage. I guess it was because I haven't tried something like this before." -P3*

7.4.3 Usability score

RQ14: How is the usability of the interfaces and what can be improved?

The SUS scores given by the three participants respectively were: 92.5, 90, 80, with an average of 87.5.

The average score got an A, top 10% of score, which means the product is most likely to be recommended to their friends.(Sauro, 2011)

As there were not many samples, we

looked at all of the three single scores and their average, which all fell into the "acceptable" zone of the SUS evaluation system, along the acceptable dimension, which means the product usability is good enough to be put to use.(Sauro, 2018)

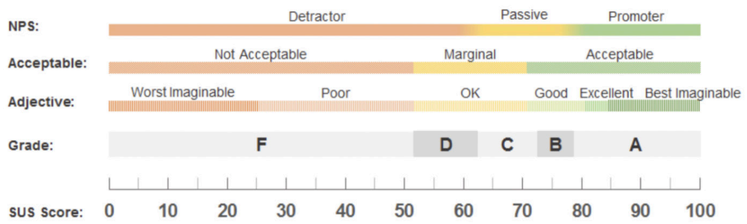


Figure 7-13. Five dimensions to interpret a SUS score

7.4.4 Recommendations for improving usability

Based on the observation and participants' feedback, here are the recommendations for improving usability:

- On the guideline/intro page, provide a full example of the coming tasks. Participants were confused and a bit nervous at the guideline/intro page because they didn't know what to expect in the following tests.

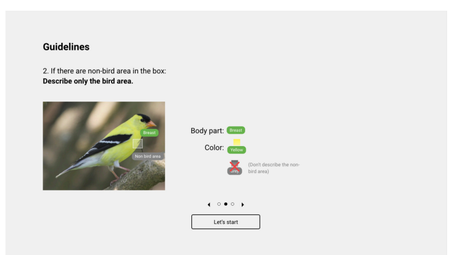


Figure 7-14. The current guideline page

- Enable guidelines to be seen on each page. All the participants didn't quite understand the guidelines and forgot about them in the tasks.
- Show a topography diagram along with/before the intro session. Participants felt confused when coming across the professional terms for the first time.
- Make the bounding box button and page indicator stand out more. They were ignored by most participants.

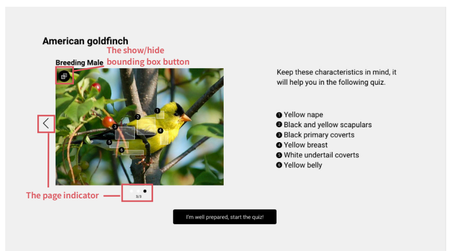


Figure 7-15. The buttons that were often neglected during the tests

- Allow room for more colors in the description, because people have different perception of colors and the current color description in the introduction session sometimes causes confusion.

"Isn't it yellow/grey/...?" (P1, P2, P3)

- Confusion on how to annotate the pattern. For both participants 2&3, it makes more sense to describe only the pattern but not the grounding colors. While all of them have forgotten what's said in the guidelines (describe both the pattern and the grounding colors).

"It should say 'select as many color as possible' ".-P1

"Because it says the color of the 'pattern', I think it should be white." -P2

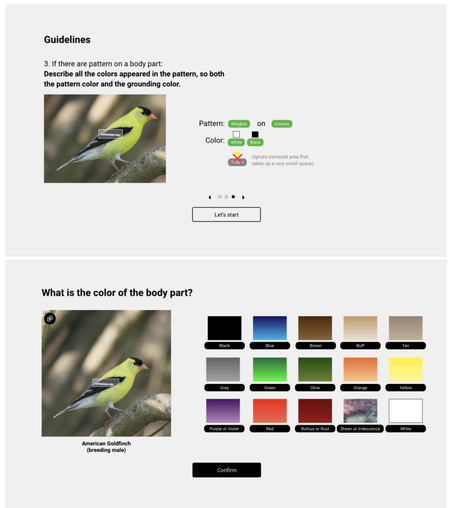


Figure 7-16. The guideline and annotation task that confused the participants

- Bring more focus to the identification of different gender/ages.

"It's still not super clear what are the differences between male/female, mature/immature birds." -P1

"It will be even funnier if there are tests of telling apart male/female birds."-P2

- Transition sentences/pages between different tasks.

"It feels a bit abrupt when suddenly a classification task comes out after the annotation task. Maybe a transition sentence like 'now you are going to learn about how to identify female goldfinch' will do better." -P3

7.4.5 Positive feedback

The participants also thought positively of some of the design ideas.

- The repetition of the annotation tasks helps to solidify the knowledge.

"The repeating has really helped to learn, it has helped me memorize names of different body parts better."-P1

- The annotation tasks help bring attention to details.

"The detailed features of birds' body parts have been marked out, which is very comprehensive."-P3

- The knowledge provided was getting deeper.

"I like how the order of the tasks was designed. It first showed photos of breeding males, which were easiest to recognize. Then it got into photos of female goldfinches, and practices on that. It is getting deeper."-P2

- The confirmation after each task was cheering, making the participant feel that they achieved something.

"I like the confirmation after each task, it makes me feel that I have achieved something."-P1

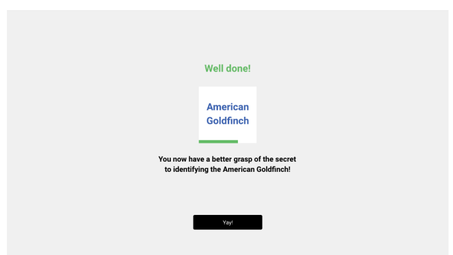


Figure 7-17. The confirmation page at the end of a learning session

Summing up Chapter 7

The aim of this chapter was to test among the target users the interfaces and the notion of gamifying the annotation process. The main goal of this chapter was to find out how they like the idea of the product in general, and what usability problems there were of the product's interactions and interfaces.

In the evaluation, it was found that the overall idea seemed attractive to most of them. The usability of the prototype was good, though some recommendations were gained for improvements.

Takeaways

- The prototype got a good score (87.5 on average) in usability according to the SUS (system usability scale).
- Overall, the interaction designed for the annotation task is easy to understand by the participants, though some details needed improving in the guidance.
- Two out of three participants thought the prototype was engaging to them, the other one didn't think so but still thought the process was fun.
- Participants thought positively about learning to identify birds with the designed annotation process.

CHAPTER 8.
FINAL DISCUSSIONS

In this chapter, insights from all the previous stages were wrapped up. To finalize, it discusses the main outcomes, answers to the research questions, design implications, future opportunities and reflects on the limitations of the project assignments.

8.1 Project summary and outcomes

In this graduation project, I carried out research around how to utilize a human-in-the-loop method for developing the explainability of machine learning models. To summarize, the following ingredients are required to enable a human-in-the-loop development process for an explainable machine learning model: citizen scientists who may be interested, strategies to motivate their use, and user-friendly interfaces. And during the project's research and design phases, these various parts of the puzzle were pieced together one by one.

In the beginning, we have roughly decided that hobbyists in birding will be the group of people we are going to design for. We observed that bird ID apps are not only a tool for identifying birds but also an important component of their learning tour of the birding field, based on interviews and online surveys with birding hobbyists (Chapters 3&4).

Then we presented birders of different levels with a few quick mock-ups, each representing part of the bird species identification model explainability, one was the representation of explanatory information, the other was the process of making annotations.

First were the explanation prototypes (Chapter5), where we learned the preference of the birding community in seeing different types of explanations, as well as validated the assumption that they can learn about bird ID with those explanations.

Then came the mock-ups of the annotation process (Chapter6), where we found out how likely it is for a general citizen to make annotations correctly on the materials coming out from ML models.

Finally (Chapter7), we put together the explanation and annotation part to develop the interfaces of a bird ID learning platform, which was tested among the users and gained the feedback of engagingness, fun, and enabling learning.

The design research conducted in this project proved the possibility of involving citizen scientists in the development of explainability for ML bird species identification models. It also demonstrated that the SECA framework can be used by ordinary citizens to collect useful annotations for developers, which opens up exciting new possibilities for future research.

8.2 Answers to the research questions

Before answering the research questions we posed, we'd like to remind readers of how these questions arose.

Initially, we wanted to find a way to engage end-users in the process, and we hoped explainability could benefit them in either learning domain knowledge or justifying the bird ID predictions. And it was discovered during the research activities that the learning aspect is more important to the birding community than the justification aspect. And, in particular, the information they can gain from a product enabled by the ML explainability is how to identify birds based on their looks.

As a result, it has become the project's design goal to enable end-users to learn to identify birds, as a way to incentivize their participation in the explainability development. And we had the following research questions around this:

Can the end-users learn about bird species identification with ML explanations and the annotation tasks?

According to the 6 user tests of the explanation prototypes conducted in Chapter5, several ways of presenting explanations were found helpful in teaching the users knowledge in bird ID. For example, by showing comparisons of two relevant species and by highlighting the contrast between them, people know better what characteristics they should pay attention to when distinguishing the species. And in the testing in Chapter 7, 3 participants indicated that the annotation tasks helped them get familiar with the features of birds.

End-users were able to learn about bird ID in a more systematic way when the explanation and annotation were combined in a game-like procedure, as we did in Chapter7. For example, participants appreciated that not only the visual traits of one species were taught, but also those of different ages and genders within that species.

Are the end-users able to make annotations needed by the developers through a game-like process?

In the testing on the annotation process in Chapter 6, 16 participants completed the 4 annotation tasks with 20 inquiries, with the help of a reference diagram on body parts, obtaining an average accuracy of more than 93%. The average accuracy is slightly lower (92%) for photographs that the classification algorithm has found more difficult to identify.

Furthermore, we discovered that all of the most popular descriptions were correct, implying that if we select all of the participants' popular descriptions, we are quite likely to receive the correct ones.

How can we engage the end-users in making annotations with a game-like bird identification learning product?

The explanations and annotation tasks were combined in a game-like process in Chapter 7, and three participants from the entrance level were examined. Overall, they agreed that using the product was an enjoyable and interesting experience. The belief that they can learn about bird ID with this process has made the product sound attractive enough to them. In addition, they also believed that the gamification element, like the textual confirmation, and varying levels of difficulty added to the process's appeal.

8.3 Design implications

In general, the result of this project suggests the possibility to recruit bird hobbyists as annotators in the loop of developing explainability for ML bird ID models.

Through the research conducted in the project, we validated both the capability of general citizens to make correct annotations on the provided materials, as well as their willingness to do so.

8.3.1 Annotation collection

Findings from this project's research (Chapters 6&7) suggest that annotation collection can be achieved through showing diagrams with professional

vocabulary on attributes for people to select from. According to our online annotation test among 16 participants, users were able to identify the proper descriptive words with an average accuracy of more than 90% using a set-up that included displaying reference diagrams on avian body parts and providing testing photographs with bounding boxes for them to annotate.

Our research also showed that the most popular descriptions collected were mostly right (100% correct in our test).

Besides, we got following information from the SECA developers:

1) They estimated that 300 sample pictures are needed for classification tasks for photos with 2 or 3 labels to provide high accuracy, high precision and high recall results, which allow to accurately explain ML models.

For additional labels, the number of sample photographs will augment. However, how it will augment depends on the intricacy of the added labels in relation to the ones that are currently there.

2) Typically, the developers choose the number of annotators per task/sample to be 3, to balance the cost and the accuracy of the annotation. Because 3 is the smallest odd number that allows us to check for annotator disagreement and collect a sufficient number of annotations. As a result, we'll need at least three different annotators for each image.

In our case, we're not sure how many example photos we'll need. It actually relies a lot on the task's complexity, such as whether the birds of different labels are difficult to distinguish or quite similar (e.g. similar shapes, colors, background, etc.). If we assume the number is n, then the total number of annotated photos required is n*3.

What's more, the findings provide the following insights for future work in annotation collection:

1) The data from Chapter 6 indicates that annotating on body parts is easier for people than on the colors. And as colors are not as important an element as bird body parts in learning, it is recommended to leave enough grey space for the recognition of colors.

2) Practices and tests in advance are necessary to ensure the high accuracy of collected annotations.

8.3.2 Designing user-friendly interactions and interfaces

From the user tests done in Chapter6 and Chapter7, we have gained the following insights on how to further enhance the user experience with the design of interactions and user interfaces:

1) Show clear guidelines on tasks, for example, by showing example and providing practices. Our research showed that participants might make false annotations because of the fuzziness in guidance.

2) The transition pages between tasks will align the experience, remind the audience of where they are, and therefore provide a better learning experience.

3) The order in which the information is delivered is crucial. People can learn about bird ID in a step-by-step manner with carefully-assign-sequence, giving them a sense of accomplishment.

8.3.3 Getting real people to use

While our study proved the feasibility of such an approach, we recognize that more work has to be done in terms of incentivizing people before this notion can be implemented as a large-scale citizen science project.

Additional particular user modeling, as well as more work on gamification design, is required to motivate users, for example, by carefully developing the leveling system, giving tasks of varying difficulty to people at various levels, or showing them with bird species that are present near their locations.

8.3.4 Future work

We focused on visual-based bird identification learning within the scope of our project because we solely used photo classification models. However, during our research, birders expressed a strong desire to learn how to identify a bird species using the information other than its visual appearance, such as its habitat, location, and behaviors, which was not included in our study.

Acknowledging that the information beyond birds' appearance is an important aspect of bird species identification, here's how we envisioned it to be possible in our existing framework:

1) Imagine we have an identification model that incorporates all of the necessary information for bird identification, such as location, pictures, habitats, and so on.

2) Then, using reference materials, human annotators might be trained to identify any of the information and describe them, which allows them to learn about bird ID at the same time.

3) As a result, the annotations they generated might be used to assist the debugging of the model as well as generate explanations for the identification results, which could also facilitate the end-users' learning.

Moreover, other than making annotations, we can also recruit citizen scientists with training to do more advanced tasks such as classify the photos of female or male birds, using a similar framework.

To conclude, the framework we investigated in this study has the potential to be used to a broader range of applications where ML classification models are used, as indicated by the above conceptualization.

8.4 Limitations

One thing we'd like to point out to readers is that the research in this project focused on the part of the SECA approach that generates local explanations, while the part that transforms local explainability into global explainability was left out of the project's scope, where more research is needed.

It is one of the limitations of this study that there is currently no quantifiable proof that the annotation procedure improved participants' ability to distinguish birds. We tested this in Chapters 6 and 7, but the tiny sample size prevented us from making definitive judgments regarding the impacts on learning.

Moreover, compared to the one required in practice, the testing of the annotation process in this study was actually

simplified, with the step of drawing bounding boxes being deleted. As a result, the annotation tasks' accuracy does not fully reflect the real-world situation.

Furthermore, due to the length of the Master's thesis, we were unable to test users with a fully interactive game-like prototype, opting instead for a static one, making the results on the overall engagement of the annotating and learning process less convincing.

To finalize, the tests weren't conducted fully in vivo. As a result, extensive testing should be conducted in order to properly understand the quality of collected annotations and people's willingness to engage in a citizen scientific project like this.

8.5 Reflection on the design assignment

In this project, we collaborated with birdwatching enthusiasts and professionals to look for ways to apply explainable AI to the birding context. Despite its exploratory nature, the study provided insight into how to collaborate with people from other fields, including the machine learning developers and the birding enthusiasts end-users.

Throughout the project we conducted various activities such as interviews, internet surveys, and prototype testing in order to understand users' expectations, as well as gathered thoughts and suggestions from the birdwatching community.

Because this is a new issue for the researcher, there are a number of design-methodology-related hurdles to overcome along the process.

Firstly, in contrast to design projects that are closely related to people's daily lives, to conduct a design research project in the field of machine learning, such as this one, where abstract notions are involved, it is difficult to elicit people's attitudes regarding those concepts through communication. As a result, we ran into certain difficulties concerning language to use during the interviews and surveys at the early stage of this project.

Moreover, when everything except the underlying technology of what we're designing for is hazy, we chose to adopt a "reverse engineering"-style approach, in which we built prototypes based on what we already had and showed them to end-users to sensitize their thoughts. This was found particularly useful for tech-focused innovation projects like this.

To sum up, this study advocated the creation of a bird-identification learning platform as a means of involving general citizens in the development loop of machine learning bird-id models' explainability. Throughout the process, we strive to find a link and balance between the development requirements and the needs of the users.

Besides, the project also shed light on how to carry out design tasks that require specific domain knowledge and the participation of people from various backgrounds. We learned along the process that presenting them prototypes of our concepts rather than asking them text-based questions was a more successful approach of eliciting their views and thoughts, especially when professional knowledge is involved.

The findings on design outcomes and approaches from this project lay the groundwork for future research into using the human-in-the-loop framework to develop interpretability for ML classification models.

REFERENCES

Anik, A. I., & Bunt, A. (2021). Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3411764.3445736

Balayn, A., Soilis, P., Lofi, C., Yang, J., & Bozzon, A. (2021). What do You Mean? Interpreting Image Classification with Crowdsourced Concept Extraction and Analysis. 12.

Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., & Belongie, S. (2010). Visual Recognition with Humans in the Loop. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), Computer Vision – ECCV 2010 (pp. 438–451). Springer. https://doi.org/10.1007/978-3-642-15561-1_32

Brooke, J. (1996). Usability evaluation in industry, chap. SUS: a "quick and dirty" usability scale. London: Taylor and Francis.

Cohen, J. M., Fink, D., & Zuckerberg, B. (2020). Avian responses to extreme weather across functional traits and temporal scales. Global Change Biology, 26(8), 4240–4250. https://doi.org/10.1111/gcb.15133

Combining artificial intelligence and citizen science to improve wildlife surveys. (2019, March 22). Mongabay Environmental News. https://news.mongabay.com/2019/03/combining-artificial-intelligence-and-citizen-science-to-improve-wildlife-surveys/

Desmet, P. (2018). Measuring Emotion: Development and Application of an Instrument to Measure Emotional Responses to Products. In M. Blythe & A. Monk (Eds.), Funology 2 (pp. 391–404). Springer. https://doi.org/10.1007/978-3-319-68213-6_25

Dhurandhar, A., Shanmugam, K., Luss, R., & Olsen, P. (2018). Improving Simple Models with Confidence Profiles. 17.

Elgendy, M. (2020). Deep learning for vision systems. Manning Publications Co.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. Science Robotics, 4(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Harrison, S. (2021). Pandemic Bird-Watching Created a Data Boom—And a Conundrum. Wired. https://www.wired.com/story/pandemic-bird-watching-created-a-data-boom-and-a-conundrum/

INaturalist Computer Vision Explorations · iNaturalist. (n.d.). iNaturalist. Retrieved October 4, 2021, from https://www.inaturalist.org/pages/computer_vision_demo

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering, 2(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0

McIntosh, P. (2014). Birding—Fun and Science. English Teaching Forum, 52(1), 36–46.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. ArXiv Preprint ArXiv:1712.00547.

Molnar, & Christoph. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Moscovitch, M. (2019). Bird Monitoring and New Media: An Anthropological Exploration. 43.

Moss, S. (2013). A bird in the bush: A social history of birdwatching. Aurum.

Nothdurft, F., Heinroth, T., & Minker, W. (2013). The Impact of Explanation Dialogues on Human-Computer Trust. In M. Kurosu (Ed.), Human-Computer Interaction. Users and Contexts of Use (Vol. 8006, pp. 59–67). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39265-8_7

Otter, K. A., Mckenna, A., LaZerte, S. E., & Ramsay, S. M. (2020). Continent-wide Shifts in Song Dialects of White-Throated Sparrows. Current Biology, 30(16), 3231-3235.e3. https://doi.org/10.1016/j.cub.2020.05.084

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ArXiv:1602.04938 [Cs, Stat]. http://arxiv.org/abs/1602.04938

Sauro, J. (2011). Measuring Usability with the System Usability Scale (SUS) – MeasuringU. https://measuringu.com/sus/

Sauro, J. (2018). 5 Ways to Interpret a SUS Score – MeasuringU. https://measuringu.com/interpret-sus-score/

Silvertown, J. (2009). A new dawn for citizen science. Trends in Ecology & Evolution, 24(9), 467–471. https://doi.org/10.1016/j.tree.2009.03.017

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. Proceedings of the 23rd International Conference on World Wide Web, 1049–1054. https://doi.org/10.1145/2567948.2579215

Sleeswijk Visser, F., van der Lugt, R., & Stappers, P. J. (2007). Sharing User Experiences in the Product Innovation Process: Participatory Design Needs Participatory Communication. Creativity and Innovation Management, 16(1), 35–45. https://doi.org/10.1111/j.1467-8691.2007.00414.x

Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 56–67. https://doi.org/10.1145/3351095.3372870

Sonka, M., Hlavac, V., & Boyle, R. (2014). Image processing, analysis, and machine vision. Cengage Learning.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation, 142(10), 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

The Double Diamond Design Process Model. (2005). Design Council.

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., & Belongie, S. (2015). Building a bird recognition app and large scale

dataset with citizen scientists: The fine print in fine-grained dataset collection. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 595–604. https://doi.org/10.1109/CVPR.2015.7298658

Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. Methods in Ecology and Evolution, 9(11), 2216–2225. https://doi.org/10.1111/2041-210X.13075

Wiersma, Y. F. (2010). Birding 2.0: Citizen Science and Effective Monitoring in the Web 2.0 World. Avian Conservation and Ecology, 5(2), art13. https://doi.org/10.5751/ACE-00427-050213