SPECIES IDENTIFICATION USING TENSORFLOW

Ali Khalighifar

(PhD Candidate in Ecology & Evolutionary Biology)

Outline

- Introduction
- Morphology-based Identification
  - **Project 1: Automated identification of Chagas disease vectors using statistical classifiers** (Completed Project)
  - **Project 2: TensorFlow improves automated identification of Chagas disease vectors** (Completed Project)
  - **Project 3: Marshalling diverse big data streams to understand risk of tick-borne diseases in the Great Plains** (Future Project)
- Signal-based Identification
  - **Project 1: Adapting TensorFlow to improve biodiversity assessment for Philippine frog species** (Current Project)
  - **Project 2: TensorFlow helps surveillance of mosquito species using cell phone recordings of wingbeats** (Current Project)
- Conclusion

What is a Species?

- At least 26 recognized species concepts

- Ernst Mayr proposed the biological species concept as:

  "Species are groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups."

- A biological species is a group of organisms that can reproduce with one another in nature and produce fertile offspring.

http://www.birds.cornell.edu

Western meadowlarks vs. Eastern meadowlarks
(*Sturnella neglecta*) (*Sturnella magna*)

Hawaiian happy-face spider
(*Theridion grallator*)

https://evolution.berkeley.edu

con rojo (figs. 139, 161) . . . . *rubrofasciata*
Primer segmento de las antenas muy pocas veces sobrepasando considerablemente el nivel del ápice del clípeo, en general apenas alcanzándolo, o aún más corto; pronoto diferente 14

14. Primer segmento del rostro más largo que el tercero (figs. 36B; 93B; 131B; 148E; 180C, D; 188B) (en caso de duda pásase al 53) 15
Primer segmento del rostro tan largo o aún más corto que el tercero (figs. 66A; 79B; 105E; 146B; 158B; 165B) 53

15. Cuerpo y corio de los hemélitros con pelos numerosos y bien perceptibles dorsalmente (figs. 16, 93D; 142A, B) 16
Cuerpo y corio prácticamente glabros, o con pelos muy breves y esparsos 21

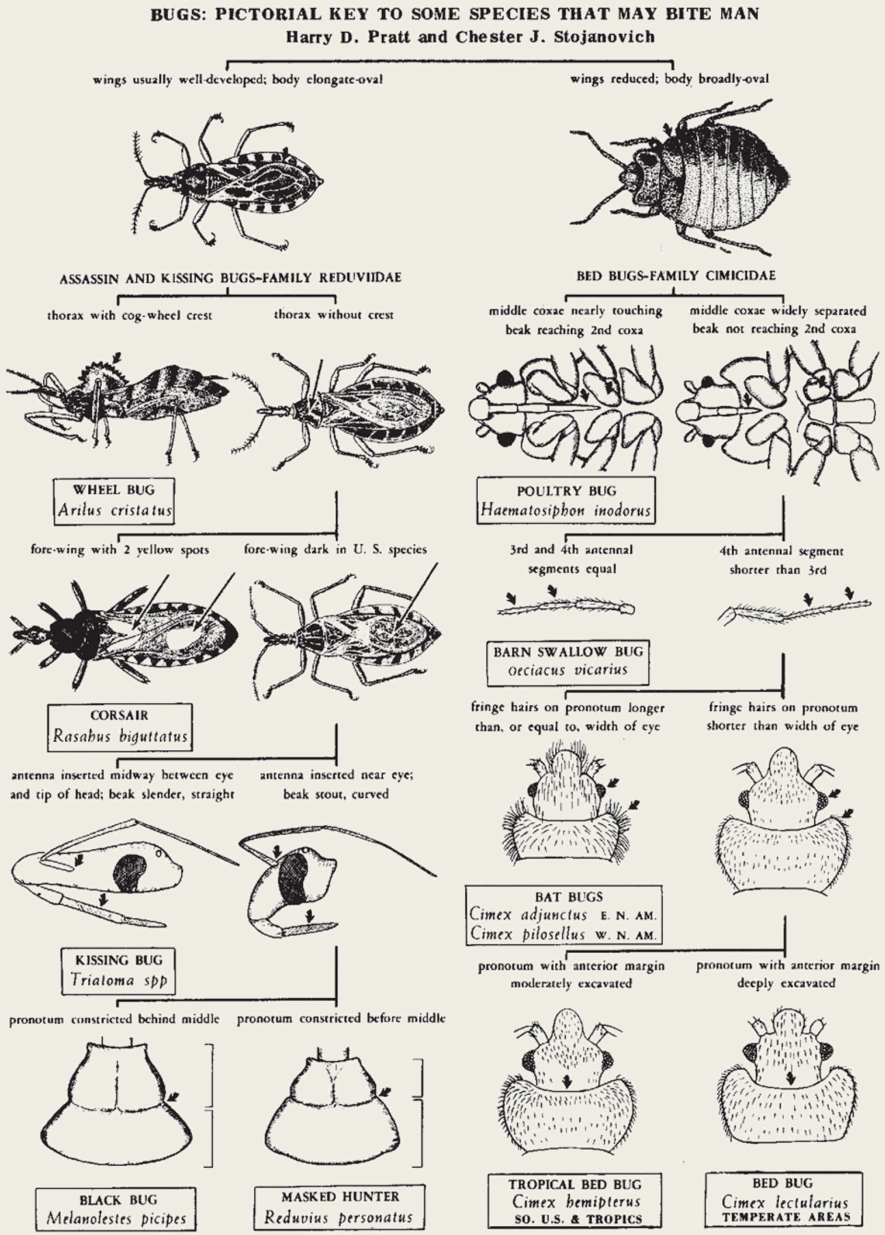
16. Cabeza fuertemente convexa dorsalmente (fig. 93B); tubérculos anteníferos alargados, relativemente próximos a los ojos (fig. 93B) . . . . *lecticularia*
Cabeza no fuertemente convexa dorsalmente (fig. 142B); tubérculos anteníferos cortos, alejados de los ojos (fig. 142B) 17

17. Corio de los hemélitros blanquecino-amarillento en su mayor parte, anaranjado en su base y negro en su ápice (fig. 130) . . *pallidipennis*
Corio sin área blanca, preponderantemente negro, solo con manchas amarillas o rojo-anaranjadas en su base y subapicalmente (figs. 101, 109, 140, 141, 143) 18

ántero-lateral (fig. 144G); raramente los segmentos del conexivo dorsal negros y con mancha amarillo-anaranjada póstero-lateral (fig. 144D-F); mesosterno, metasterno y vientre del abdómen siempre con pelos suberectos largos . . . . *picturata*
Genas con frecuencia sin atingir el nivel del ápice del clípeo (fig. 102D); pronoto con lóbulo posterior totalmente negro (fig. 101) o con 1+1 pequeñas manchas claras sobre los ángulos humerales (fig. 102D); segmentos del conexivo dorsal negros con mancha amarilla o amarillo-anaranjada en el tercio o en la mitad posterior, que se extiende o no hasta la sutura conexival (fig. 102B, C, D); mesosterno con pelos suberectos largos; metasterno y vientre del abdómen con pelos semejantes o con pelos cortos y acostados . . . . *longipennis*

21. Disco del escudete en la base con 1+1 tubérculos salientes dirigidos hacia adelante y tocando el borde posterior del pronoto (figs. 70B; 126D); ángulos humerales explanados (figs. 70B; 126D) 22
Disco del escudete sin los tubérculos mencionados, ángulos humerales no explanados 23

22. Color general castaño claro (fig. 70A); tubérculos discales y laterales del lóbulo anterior del pronoto muy salientes (fig. 70B); fémures anteriores y medianos con



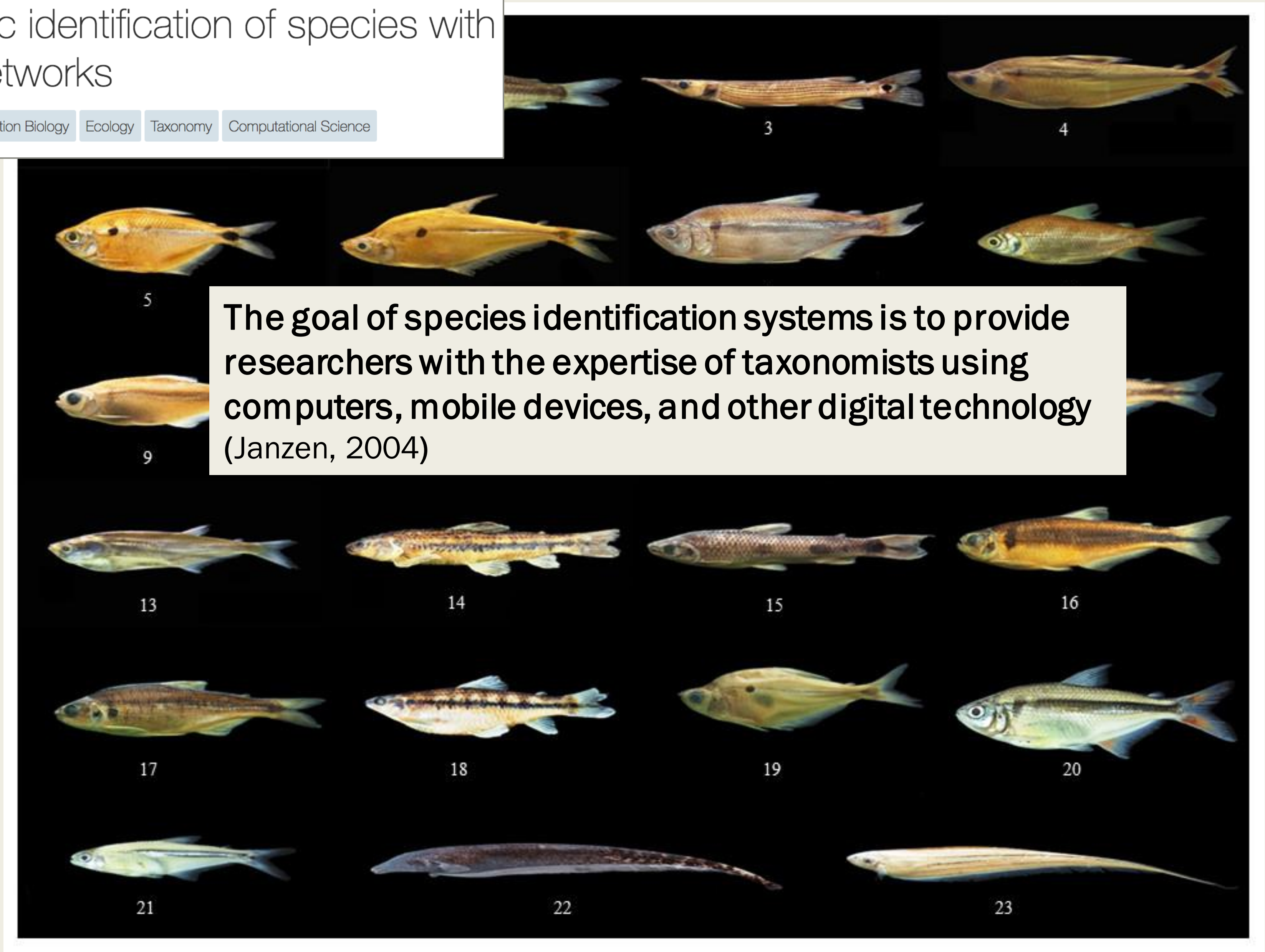
PeerJ

Related research

PEER-REVIEWED

Automatic identification of species with neural networks

Biodiversity | Conservation Biology | Ecology | Taxonomy | Computational Science



The goal of species identification systems is to provide researchers with the expertise of taxonomists using computers, mobile devices, and other digital technology (Janzen, 2004)

Outline

- Introduction
- Morphology-based Identification
  - **Project 1: Automated identification of Chagas disease vectors using statistical classifiers** (Completed Project)
  - **Project 2: TensorFlow improves automated identification of Chagas disease vectors** (Completed Project)
  - **Project 3: Marshalling diverse big data streams to understand risk of tick-borne diseases in the Great Plains** (Future Project)
- Signal-based Identification
  - **Project 1: Adapting TensorFlow to improve biodiversity assessment for Philippine frog species** (Current Project)
  - **Project 2: TensorFlow helps surveillance of mosquito species using cell phone recordings of wingbeats** (Current Project)
- Conclusion

PROJECT 1: VIRTUAL VECTOR LAB

Spencer Art Museum, ITTC, Biodiversity Institute @ KU
Universidade de Brasília
Instituto Nacional de Salud Pública de Mexico

PeerJ

Automated identification of insect vectors of Chagas disease in Brazil and Mexico: the Virtual Vector Lab

Rodrigo Gurgel-Gonçalves[1], Ed Komp[2], Lindsay P. Campbell[3], Ali Khalighifar[3], Jarrett Mellenbruch[4], Vagner José Mendonça[1,5], Hannah L. Owens[3,6], Keynes de la Cruz Felix[7], A Townsend Peterson[3] and Janine M. Ramsey[7]

[1] Faculty of Medicine, Universidade de Brasília, Brasilia, DF, Brazil
[2] Information and Telecommunication Technology Center, University of Kansas, Lawrence, KS, United States
[3] Biodiversity Institute, University of Kansas, Lawrence, KS, United States
[4] Spencer Art Museum, University of Kansas, Lawrence, KS, United States
[5] Centro de Ciências da Saúde, Universidade Federal do Piauí, Brazil
[6] Florida Museum of Natural History, University of Florida, Gainesville, FL, United States
[7] Centro Regional de Investigación en Salud Pública, Instituto Nacional de Salud Publica, Tapachula, Chiapas, Mexico

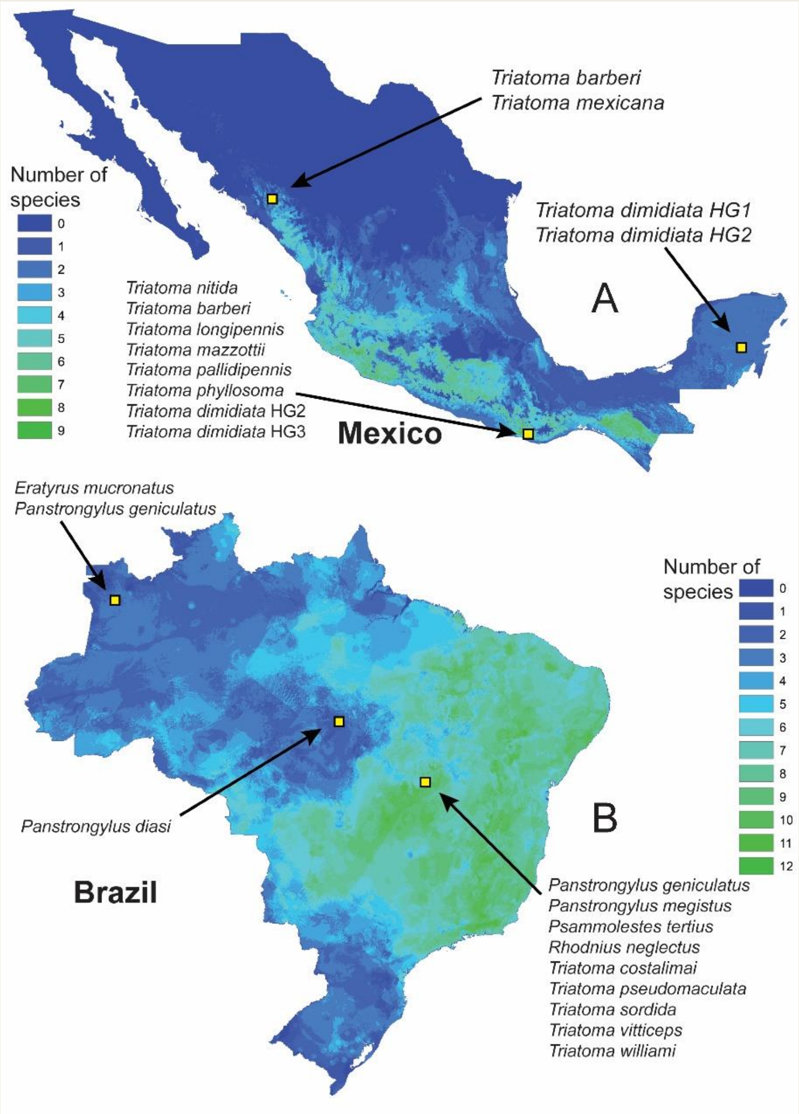
ABSTRACT

Identification of arthropods important in disease transmission is a crucial, yet difficult, task that can demand considerable training and experience. An important case in point is that of the 150+ species of Triatominae, vectors of *Trypanosoma cruzi*, causative agent of Chagas disease across the Americas. We present a fully automated system that is able to identify triatomine bugs from Mexico and Brazil with an accuracy consistently above 80%, and with considerable potential for further improvement. The system processes digital photographs from a photo apparatus into landmarks, and uses ratios of measurements among those landmarks, as well as (in a preliminary exploration) two measurements that approximate aspects of coloration, as the basis for classification. This project has thus produced a working prototype that achieves reasonably robust correct identification rates, although many more developments can and will be added, and—more broadly—the project illustrates the value of multidisciplinary collaborations in resolving difficult and complex challenges.

Submitted 26 September 2016
Accepted 28 January 2017
Published 18 April 2017

Corresponding author
A Townsend Peterson, town@ku.edu

Subjects Entomology, Computational Science
Keywords Identification, Chagas disease, Triatominae, Automation, Primary occurrence data

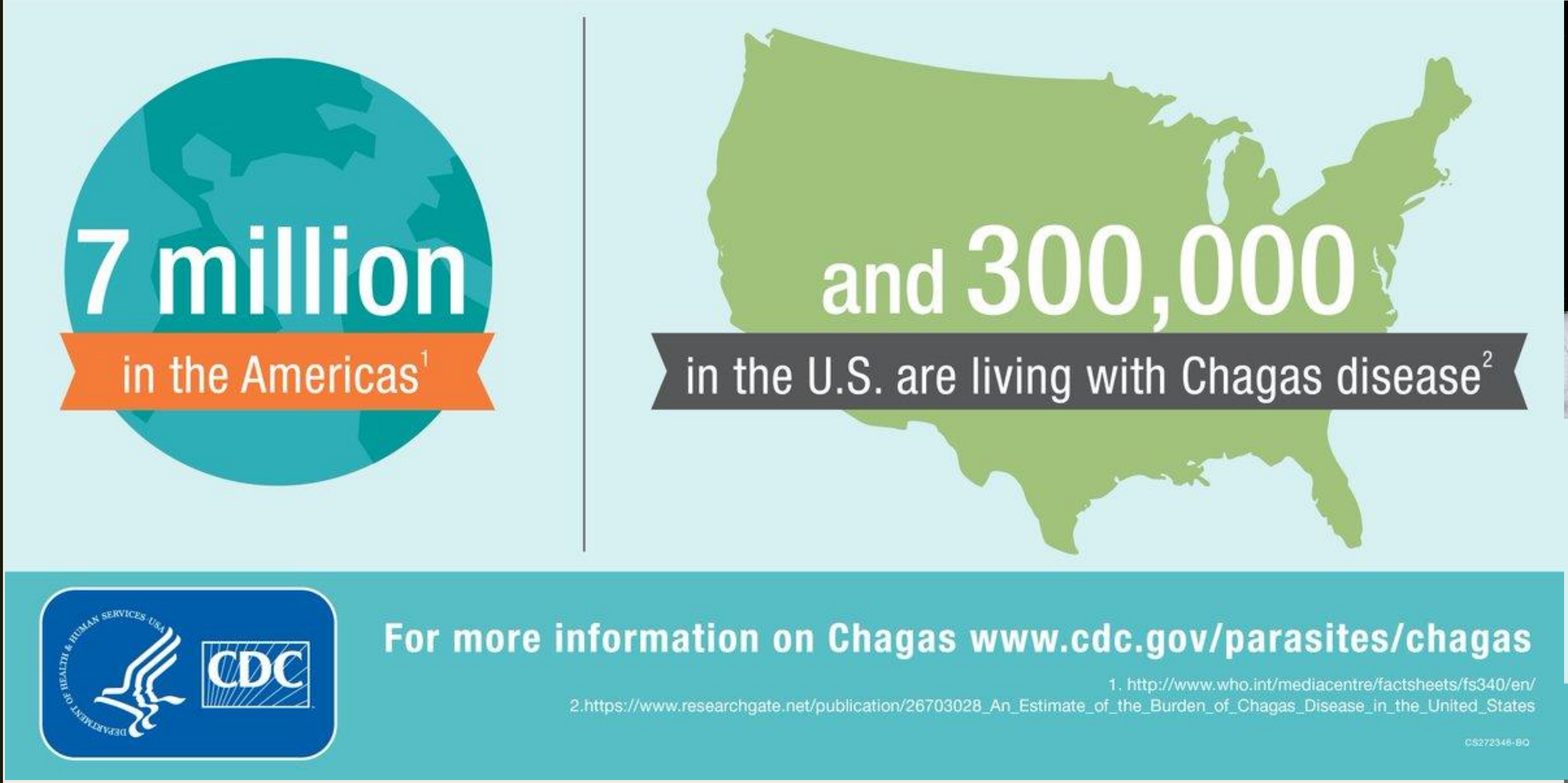


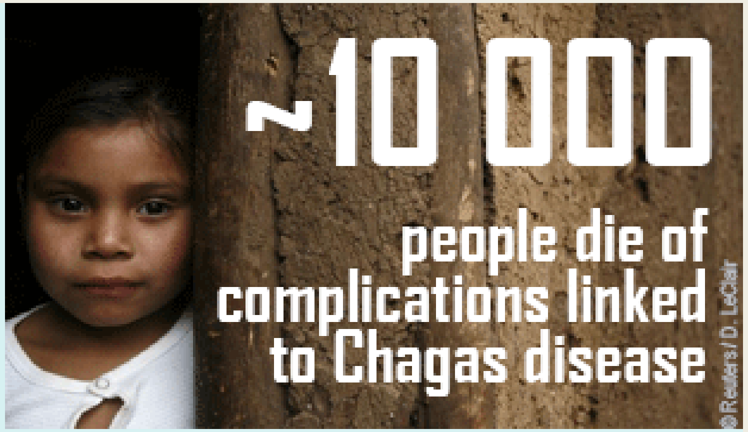
- **Vector-Borne Disease:** Disease that results from an infection transmitted to humans and other animals by *vectors* (blood-feeding anthropods)

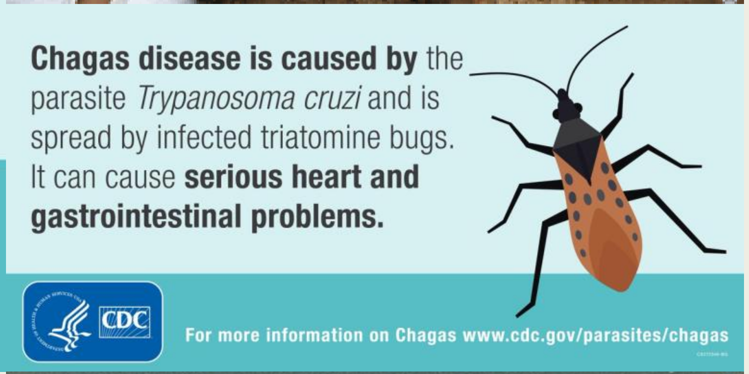
  - **Vector-borne diseases account for more than 17% of all infectious diseases, causing more than 700,000 deaths annually.**

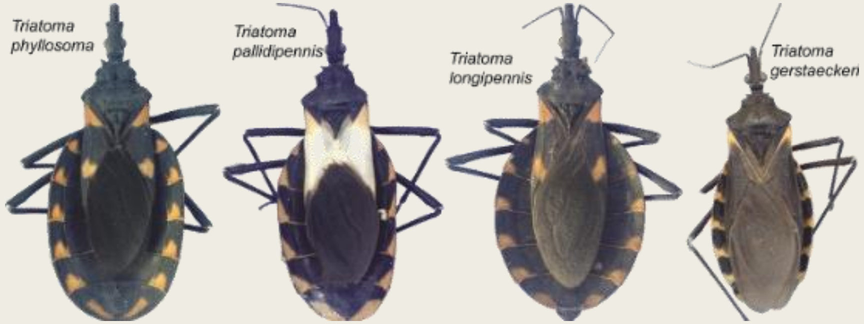
- **Vectors:** Living organisms that can transmit infectious diseases between humans or from animals to humans. Vectors ingest disease-producing microorganisms during a blood meal from an infected host (human or animal) and later inject it into a new host during their subsequent blood meal.

Chagas Disease







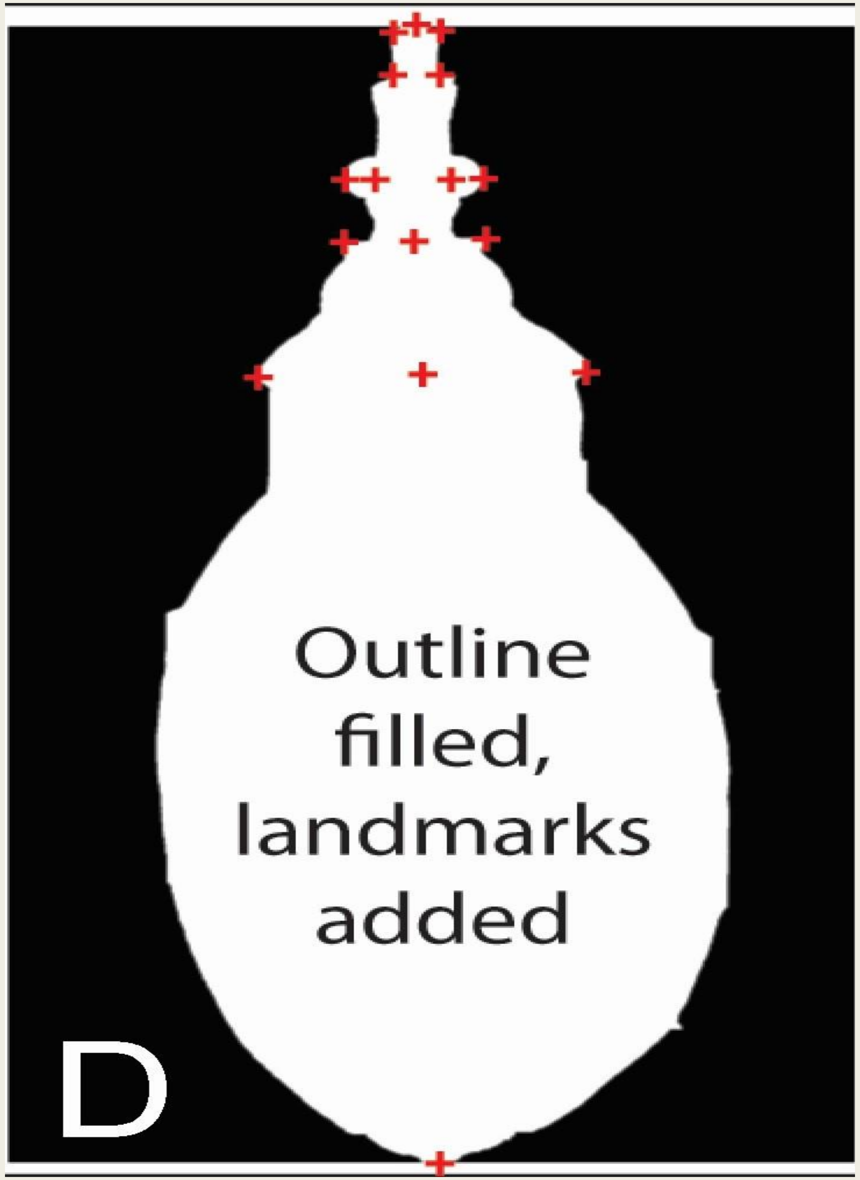


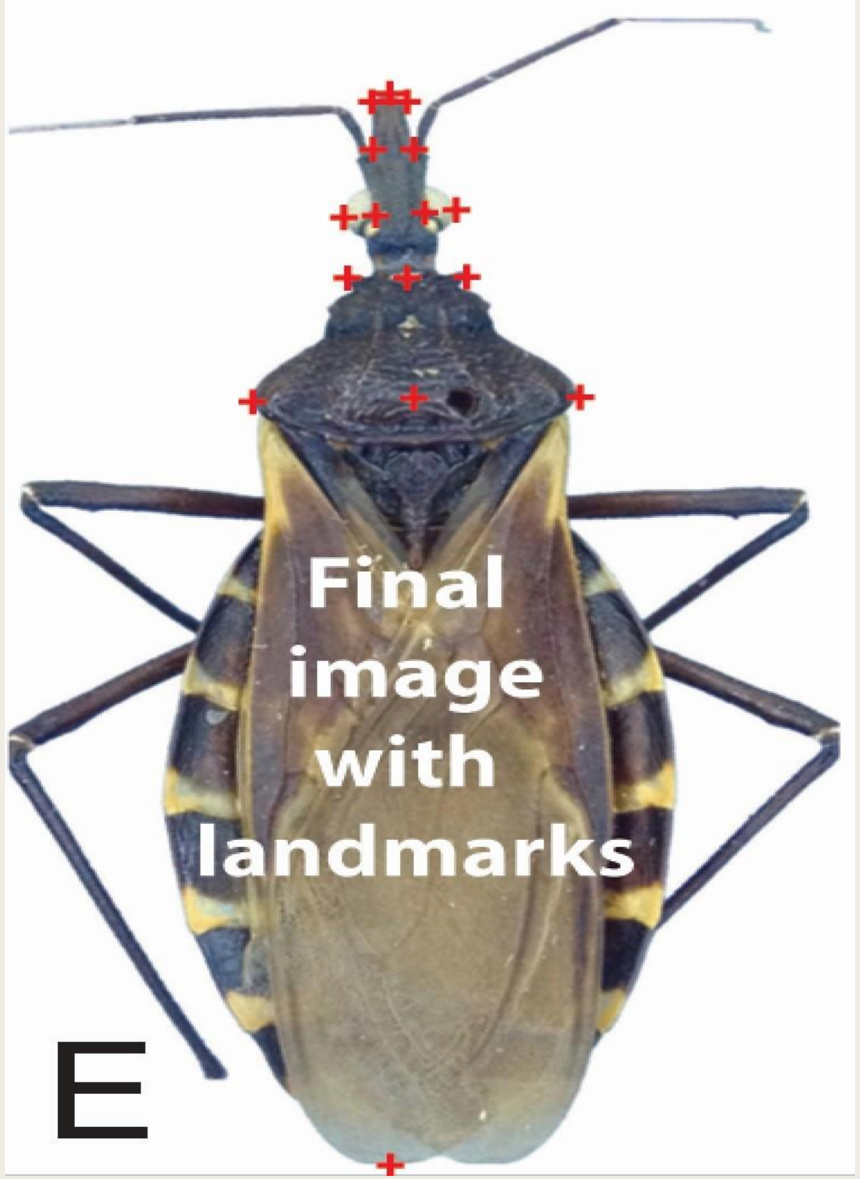

DURACELL


size
standard
Original
image
A


Binary
version
B

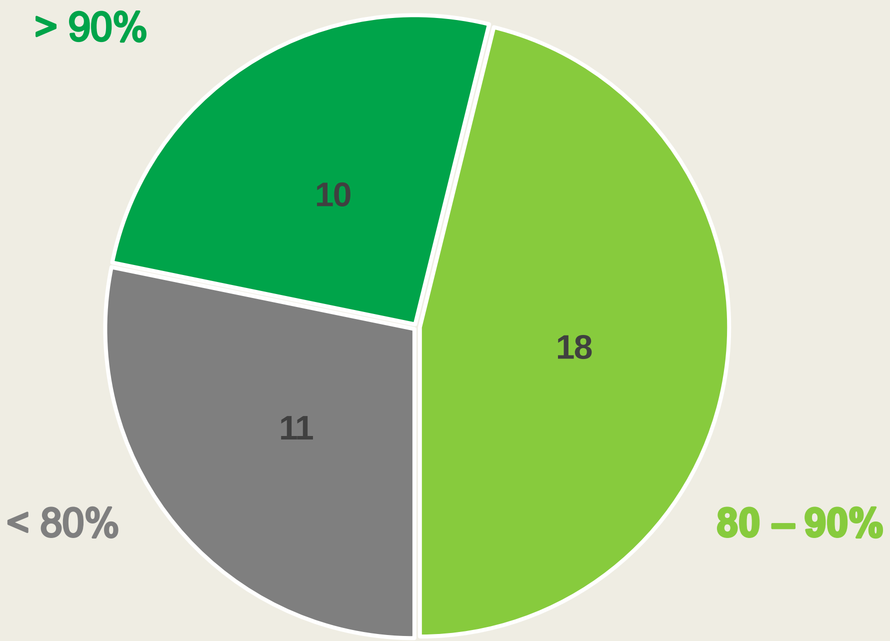

Legs and antennae removed, converted to outline
C


Outline
filled,
landmarks
added
D

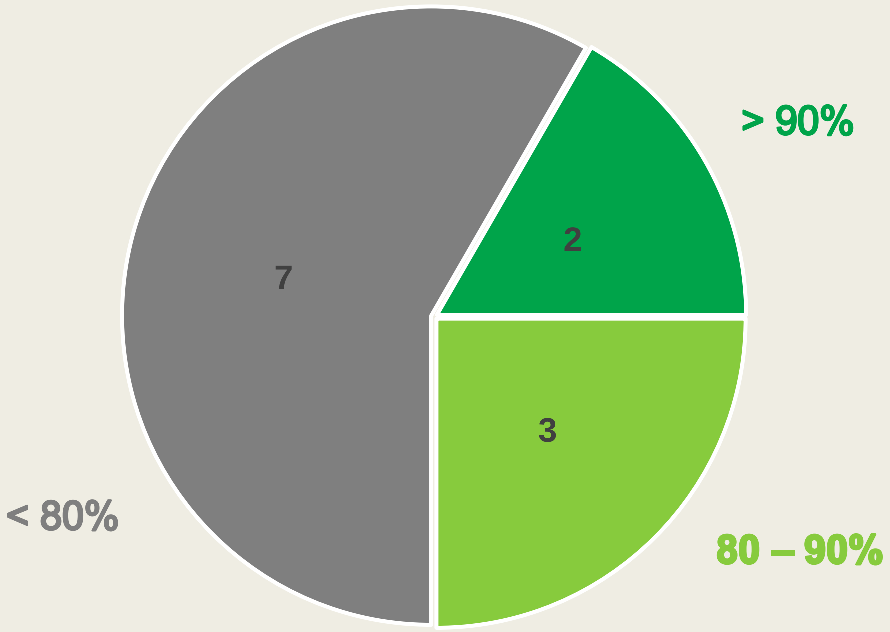

Final
image
with
landmarks
E

Results

39 Brazilian Species → Overall Rate: 83.3%



12 Mexican Species → Overall Rate: 74.9%



Outline

- Introduction
- Morphology-based Identification
  - **Project 1: Automated identification of Chagas disease vectors using statistical classifiers** (Completed Project)
  - **Project 2: TensorFlow improves automated identification of Chagas disease vectors** (Completed Project)
  - **Project 3: Marshalling diverse big data streams to understand risk of tick-borne diseases in the Great Plains** (Future Project)
- Signal-based Identification
  - **Project 1: Adapting TensorFlow to improve biodiversity assessment for Philippine frog species** (Current Project)
  - **Project 2: TensorFlow helps surveillance of mosquito species using cell phone recordings of wingbeats** (Current Project)
- Conclusion

Volume 56, Issue 5
September 2019

< Previous Next >

Deep Learning Algorithms Improve Automated Identification of Chagas Disease Vectors

Ali Khalighifar ✉, Ed Komp, Janine M Ramsey, Rodrigo Gurgel-Gonçalves, A Townsend Peterson

Journal of Medical Entomology, Volume 56, Issue 5, September 2019, Pages 1404–1410, https://doi.org/10.1093/jme/tjz065
Published: 23 May 2019 **Article history** ▾

Cite Permissions Share ▾

Abstract

Vector-borne Chagas disease is endemic to the Americas and imposes significant economic and social burdens on public health. In a previous contribution, we presented an automated identification system that was able to discriminate among 12 Mexican and 39 Brazilian triatomine (Hemiptera: Reduviidae) species from digital images. To explore the same data more deeply using machine-learning approaches, hoping for improvements in classification, we employed TensorFlow, an open-source software platform for a deep learning algorithm. We trained the algorithm based on 405 images for Mexican triatomine species and 1,584 images for Brazilian triatomine species. Our system achieved 83.0 and 86.7% correct identification rates across all Mexican and Brazilian species, respectively, an improvement over comparable rates from statistical classifiers (80.3 and 83.9%, respectively). Incorporating distributional information to reduce numbers of species in analyses improved identification rates to 95.8% for Mexican species and 98.9% for Brazilian species. Given the 'taxonomic impediment' and difficulties in providing entomological expertise necessary to control such diseases, automating the identification process offers a potential partial solution to crucial challenges.

Keywords: Chagas disease, TensorFlow, deep learning, Triatominae, automated species identification

Issue Section: Vector-Borne Diseases, Surveillance, Prevention

TensorFlow

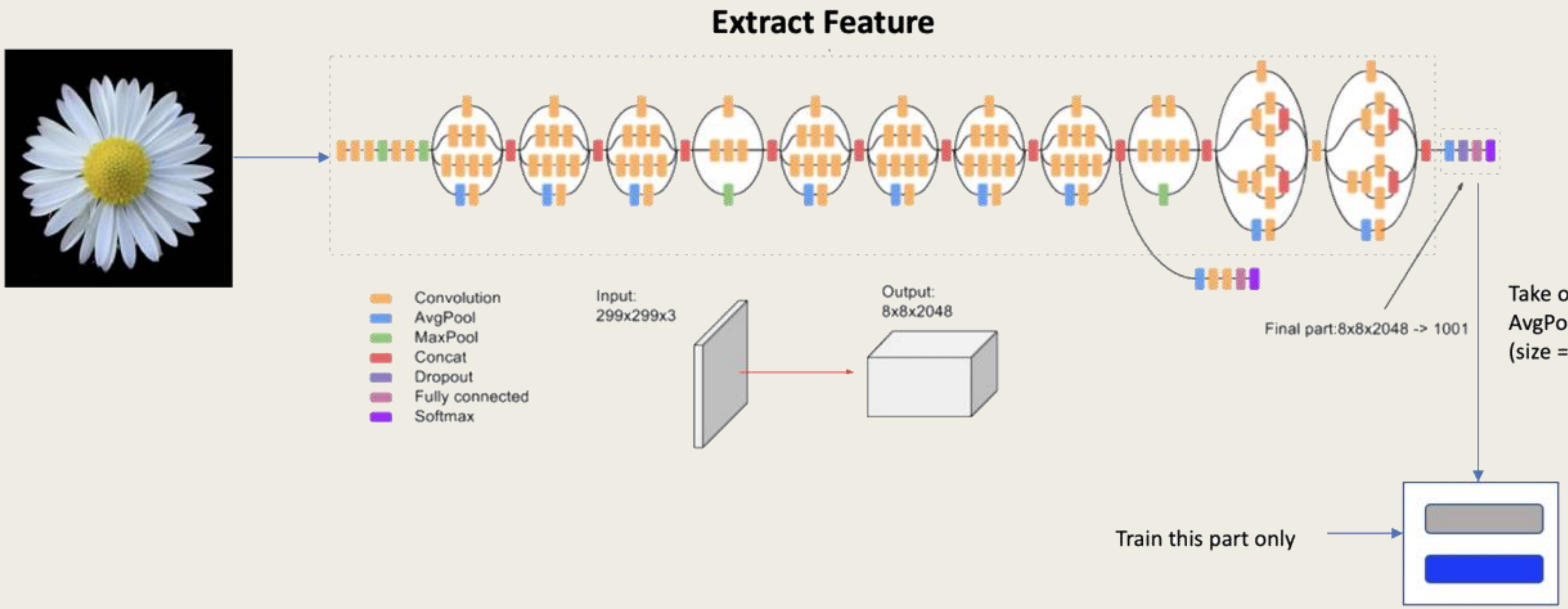
- Transfer learning using Inception v3

- Google Brain Team designed Inception v3, a convolutional neural network for ILSVRC (ImageNet Large Scale Visual Recognition Competition) that is 48 layers deep and can classify images into 1000 object categories

*"The Inception deep convolutional architecture was introduced as **GoogLeNet** in (Szegedy et al. 2015a), here named **Inception v1**. Later the Inception architecture was refined in various ways, first by the introduction of **batch normalization** (Ioffe and Szegedy 2015) (**Inception v2**). Later by additional **factorization** ideas in the third iteration (Szegedy et al. 2015b) which will be referred to as **Inception v3**."*

Methods

- ➢ Leave-one-out cross validation
- ➢ Processed images vs non-processed images
- ➢ Distributional data

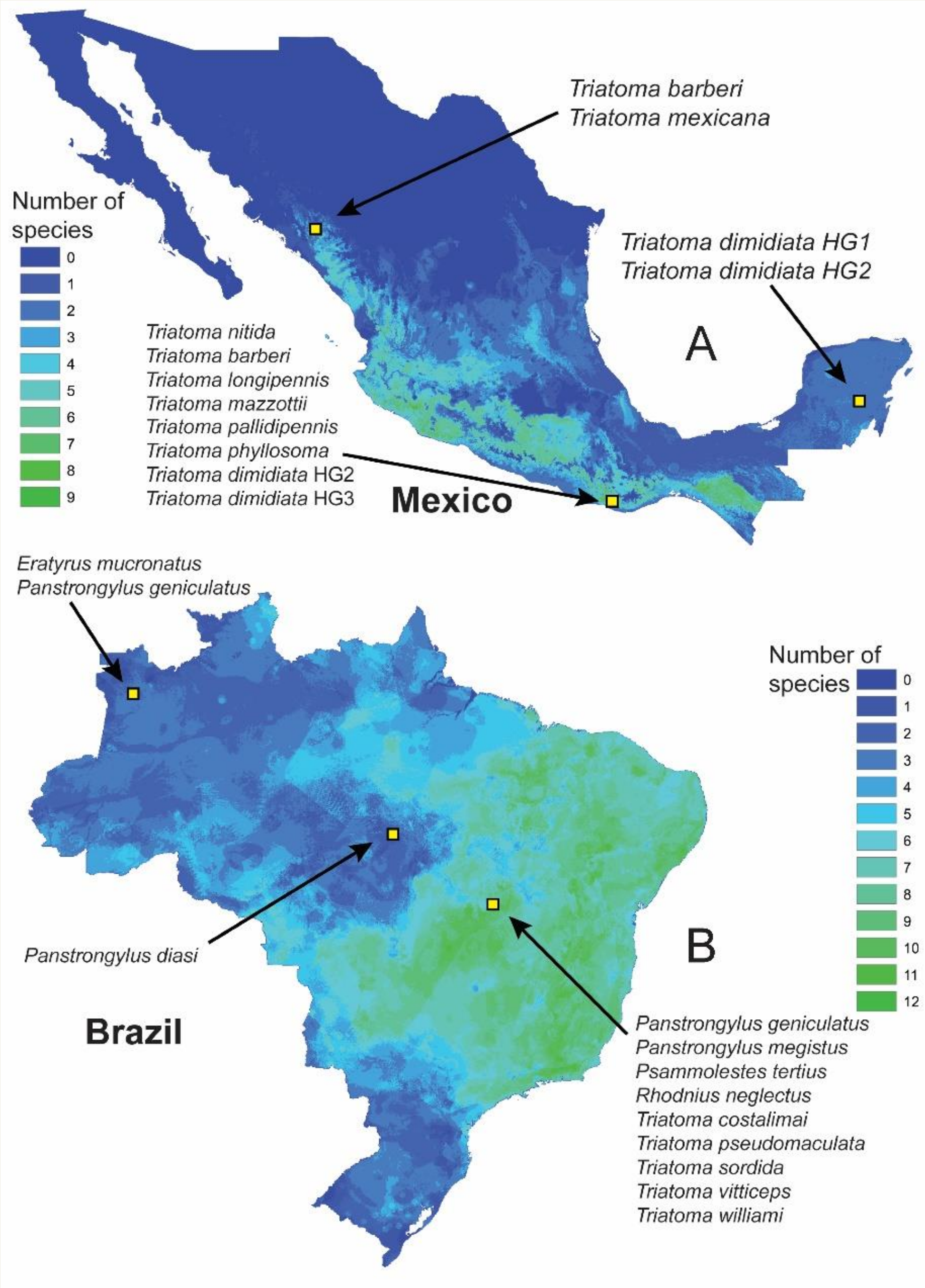


Methods

- Leave-one-out cross validation
- Processed images vs non-processed images
- Distributional data

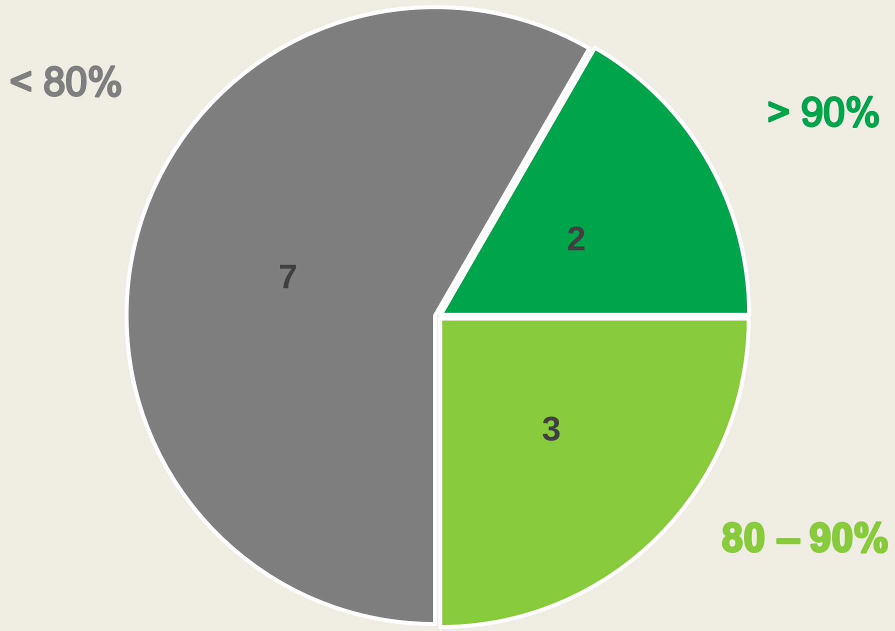
Methods

- Leave-one-out cross validation
- Processed images vs non-processed images
- Distributional data



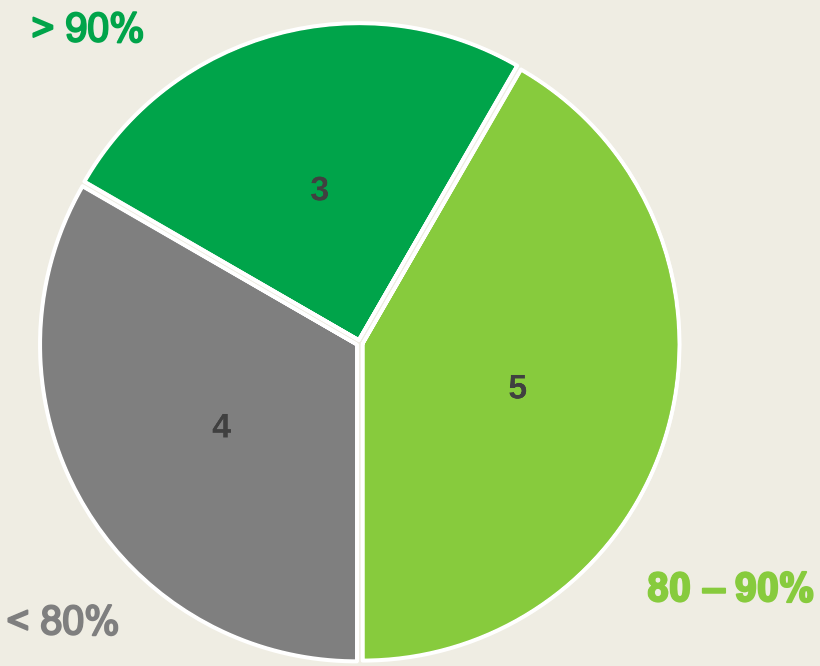
Previous Results

12 Mexican Species → Overall Rate: 74.9%



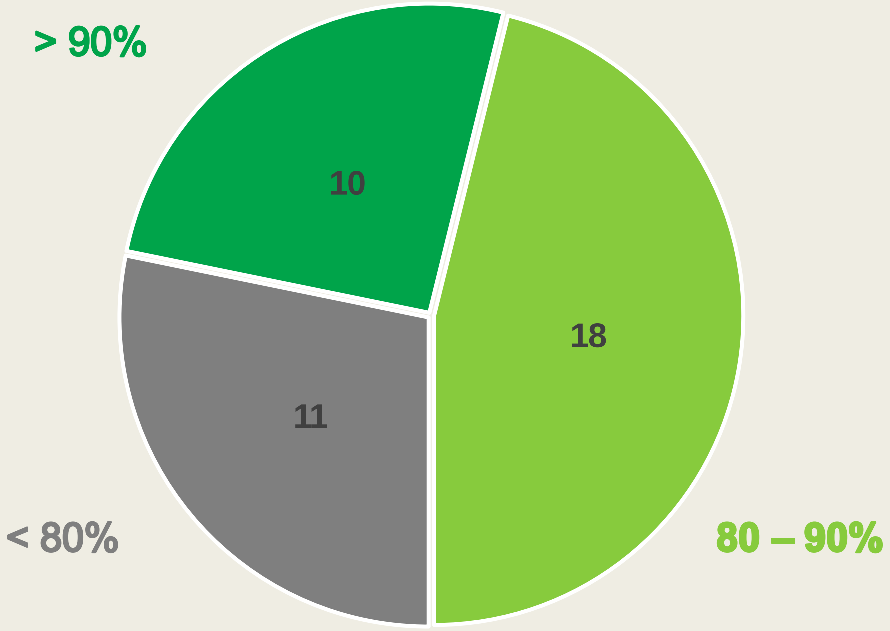
TensorFlow

12 Mexican Species → Overall Rate: 83.0%



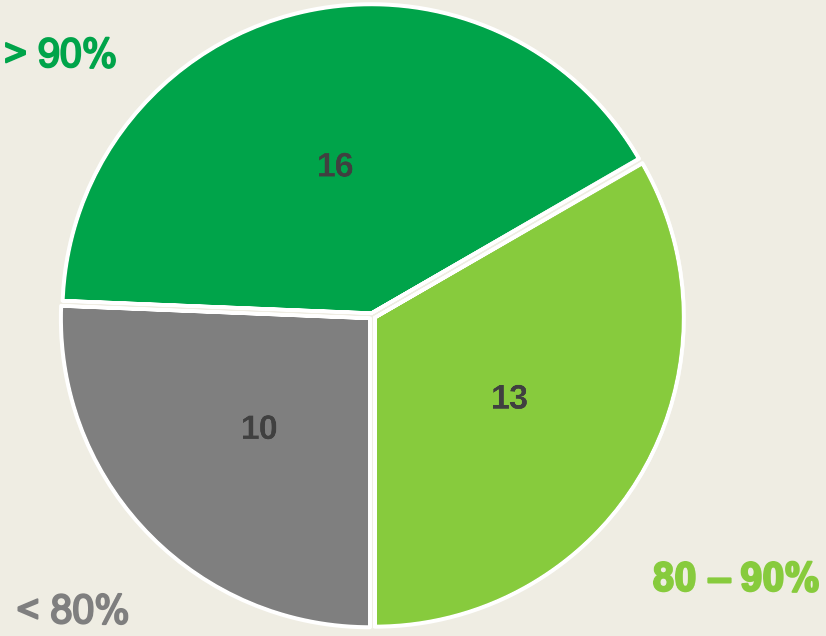
Previous Results

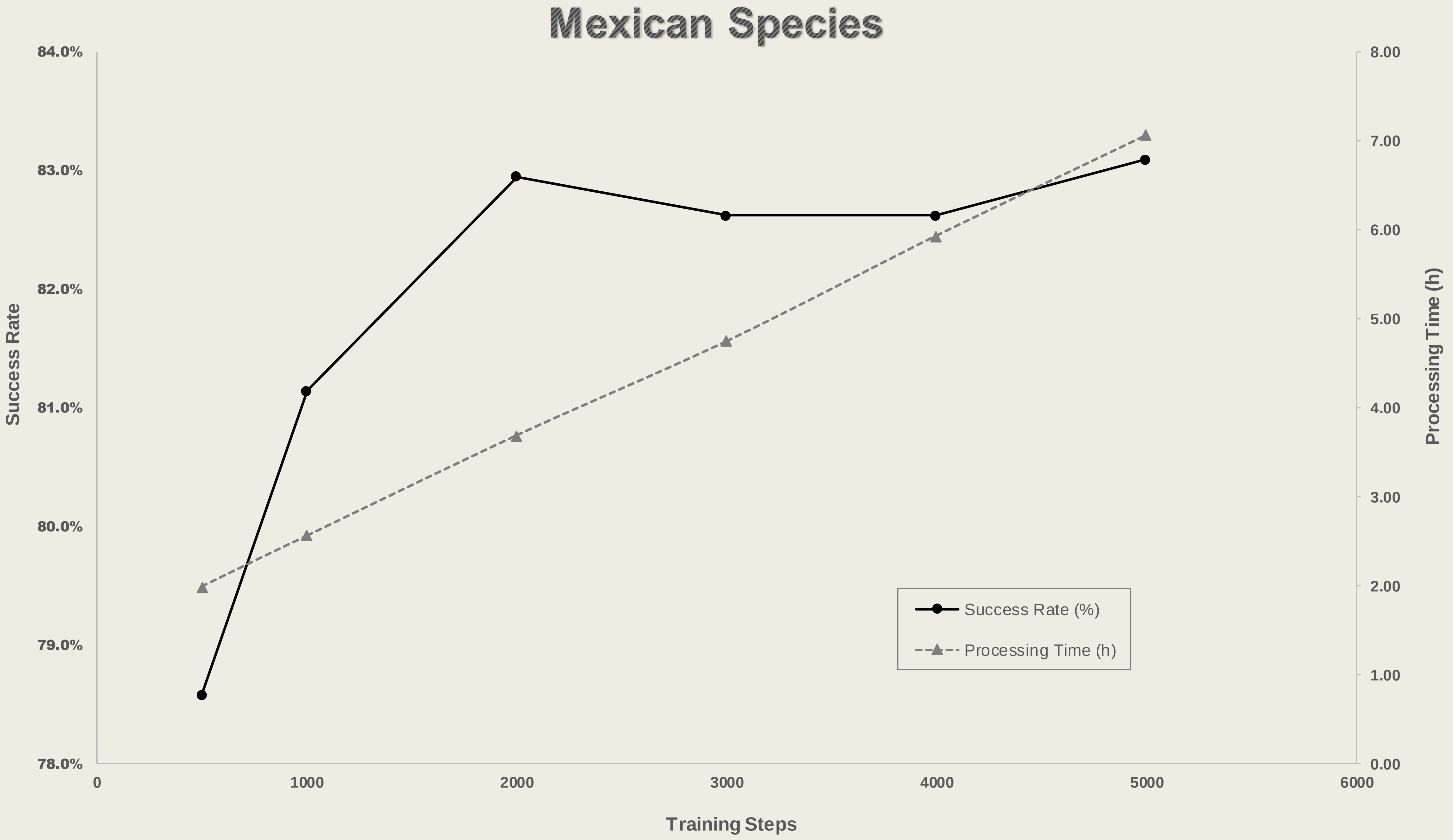
39 Brazilian Species → Overall Rate: 83.3%

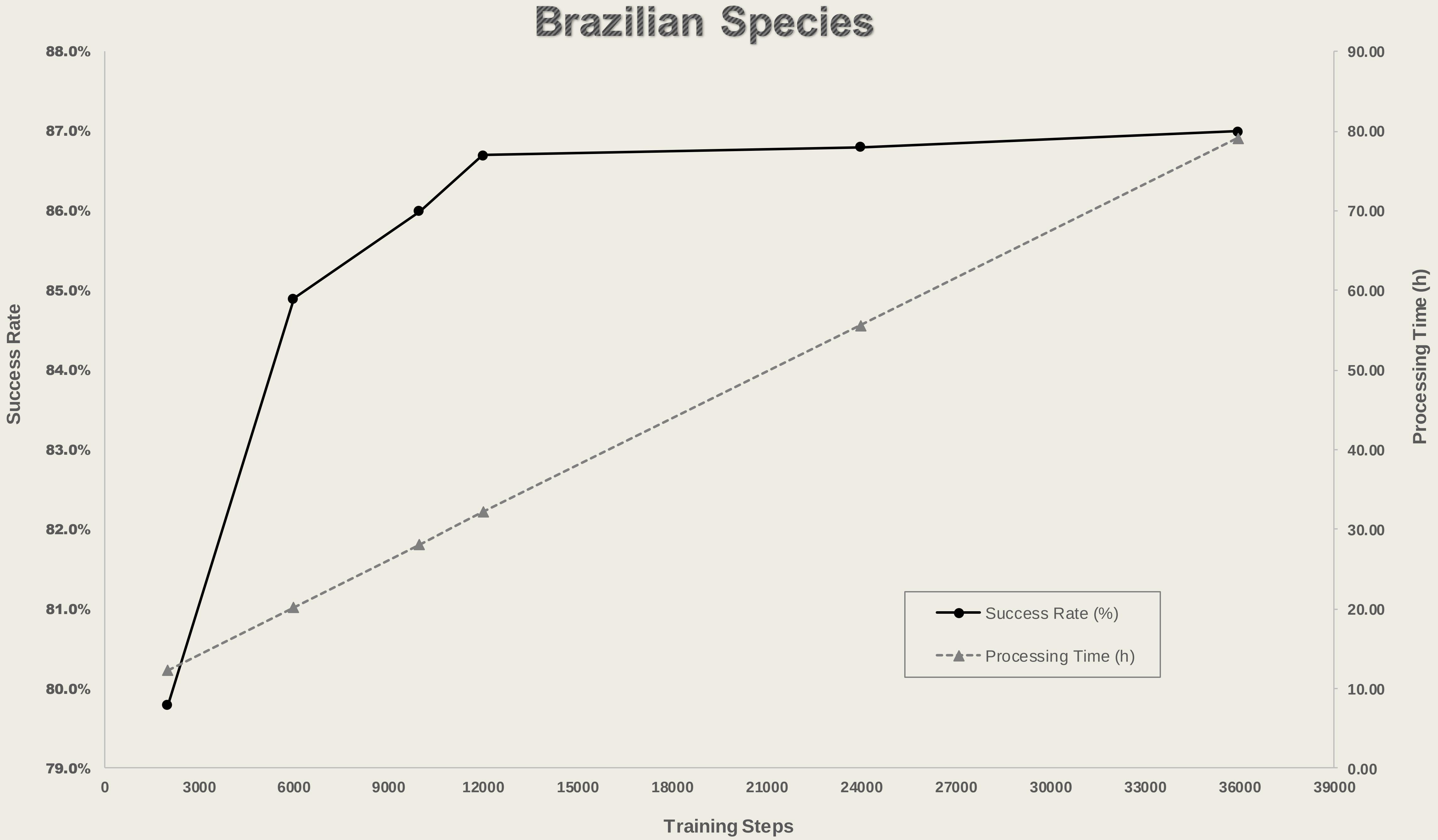


TensorFlow

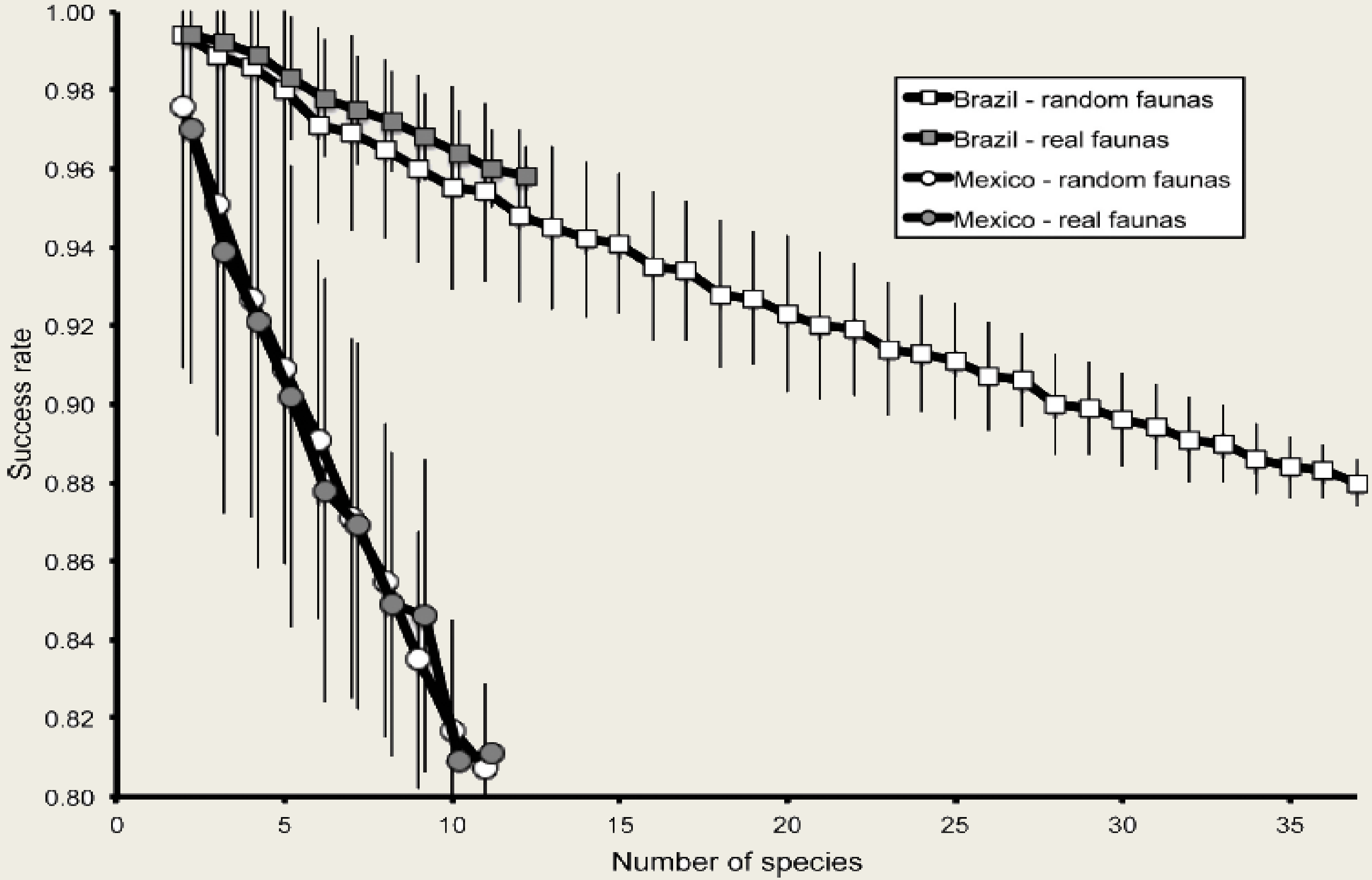
39 Brazilian Species → Overall Rate: 86.7%




Mexican Species
Success Rate
Processing Time (h)
Training Steps
84.0%
83.0%
82.0%
81.0%
80.0%
79.0%
78.0%
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
0
1000
2000
3000
4000
5000
6000
Success Rate (%)
Processing Time (h)


Brazilian Species
Success Rate
Processing Time (h)
Training Steps
88.0%
87.0%
86.0%
85.0%
84.0%
83.0%
82.0%
81.0%
80.0%
79.0%
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0
3000
6000
9000
12000
15000
18000
21000
24000
27000
30000
33000
36000
39000
Success Rate (%)
Processing Time (h)

Distributional Data

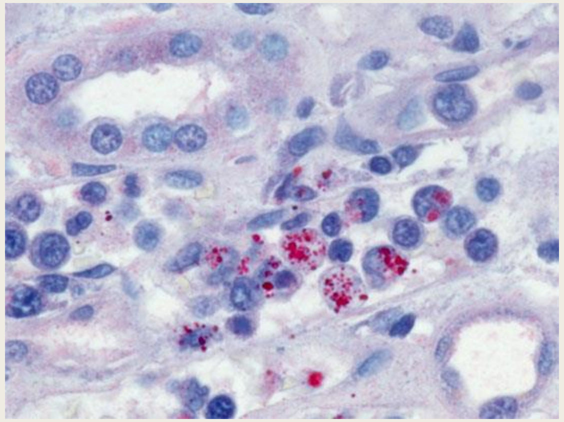


Outline

- Introduction
- Morphology-based Identification
  - **Project 1: Automated identification of Chagas disease vectors using statistical classifiers** (Completed Project)
  - **Project 2: TensorFlow improves automated identification of Chagas disease vectors** (Completed Project)
  - **Project 3: Marshalling diverse big data streams to understand risk of tick-borne diseases in the Great Plains** (Future Project)
- Signal-based Identification
  - **Project 1: Adapting TensorFlow to improve biodiversity assessment for Philippine frog species** (Current Project)
  - **Project 2: TensorFlow helps surveillance of mosquito species using cell phone recordings of wingbeats** (Current Project)
- Conclusion

Tick-borne diseases

- ❖ Lyme disease
- ❖ Ehrlichiosis
- ❖ Anaplasmosis
- ❖ Babesiosis
- ❖ Tick-borne relapsing fever
- ❖ Tularemia
- ❖ Rocky Mountain spotted fever
- ❖ Southern tick-associated rash illness



For the first time, this project marshals deployment, integration, pattern analysis and modeling of four big data streams in order to address emerging challenges of tick-borne diseases in the southern Great Plains:

- Synthesize historic and current occurrence data for tick specimens
- Generate genomic data on ticks and pathogens to identify tick species and characterize the suite of pathogens that they carry
- Gather remote sensing data to characterize the region's environmental landscapes
- Identify tick species using deep neural networks

Outline

- Introduction
- Morphology-based Identification
  - **Project 1: Automated identification of Chagas disease vectors using statistical classifiers** (Completed Project)
  - **Project 2: TensorFlow improves automated identification of Chagas disease vectors** (Completed Project)
  - **Project 3: Marshalling diverse big data streams to understand risk of tick-borne diseases in the Great Plains** (Future Project)
- Signal-based Identification
  - **Project 1: Adapting TensorFlow to improve biodiversity assessment for Philippine frog species** (Current Project)
  - **Project 2: TensorFlow helps surveillance of mosquito species using cell phone recordings of wingbeats** (Current Project)
- Conclusion

You may ask …

- ➢ Why study frogs?

- ➢ Why study the Philippine biodiversity?

- ➢ Why analyze the calls/signals?

❖ Why study frogs?

- **Play an important role in the food chain**
- **Excellent bio-indicators**
- **Control insect populations**

➢ Why study the Philippine biodiversity?

➢ Why analyze the calls/signals?

➢ Why study frogs?

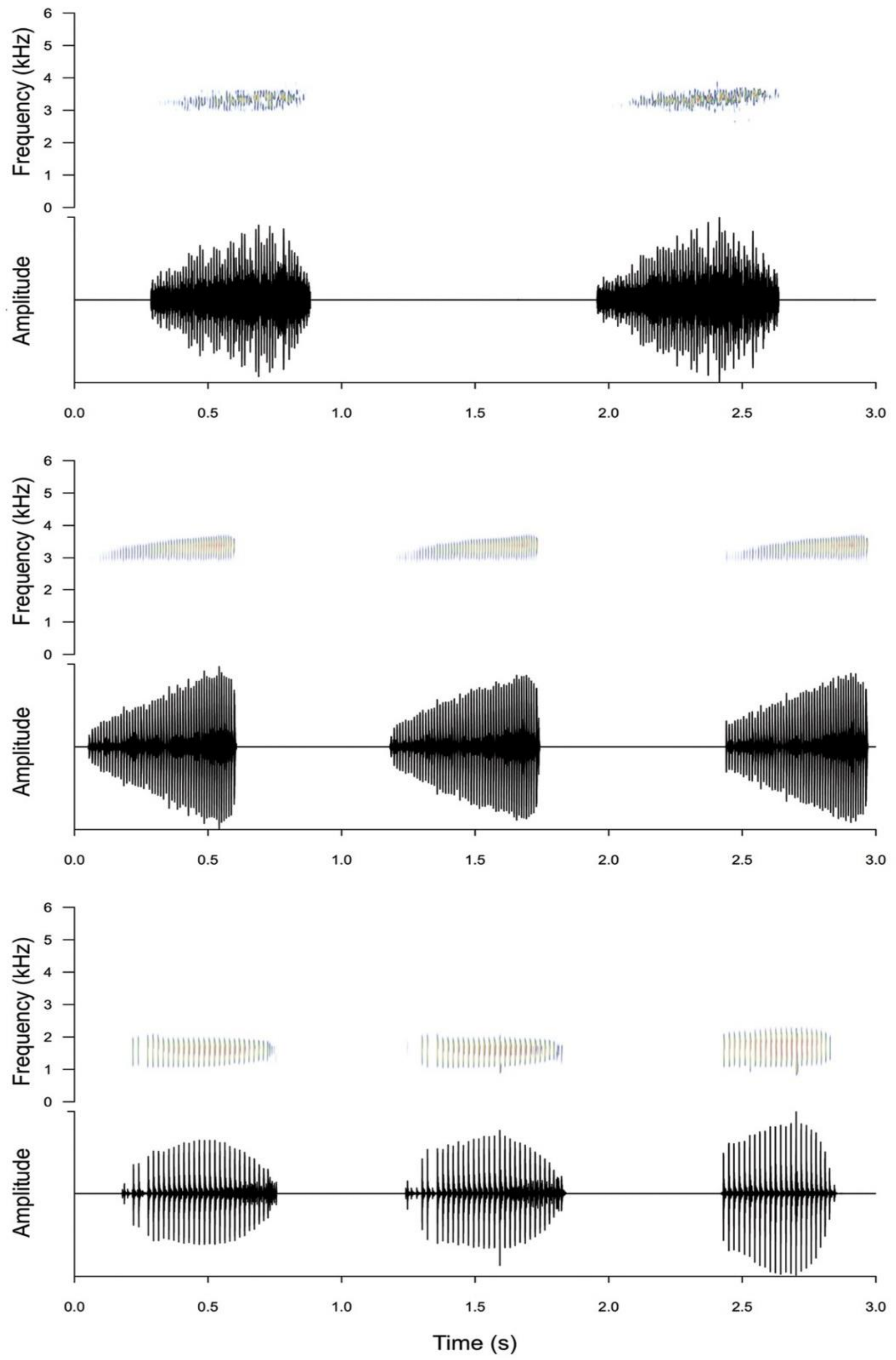
❖ **Why study the Philippine biodiversity?**

- **Biodiversity hotspot**
- **Ranked first in terms of amphibian endemism (98 out of 120 sp.)**
- **High rate of species discovery**
- **Availability of data**

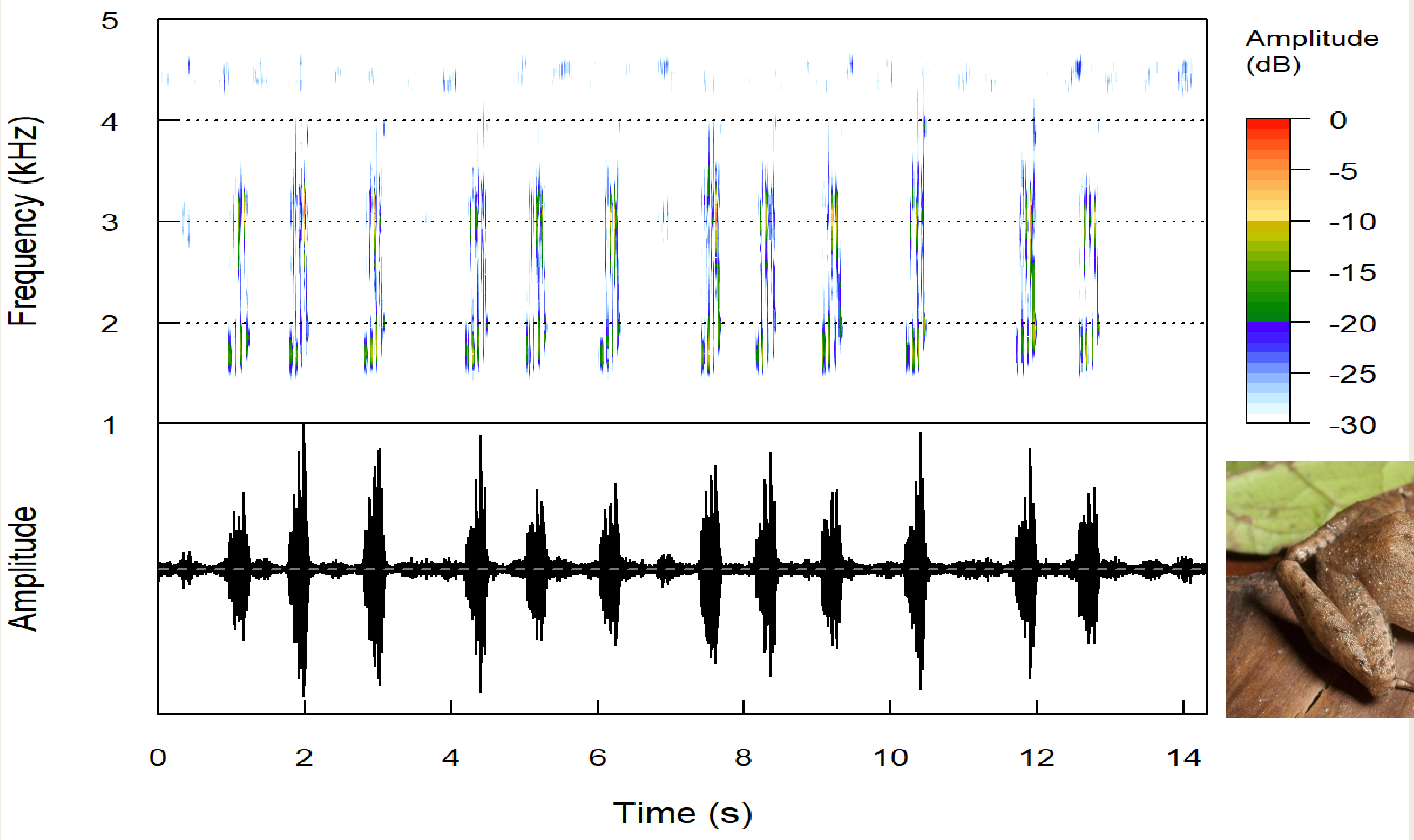
➢ Why analyze the calls/signals?

- Why study frogs?
- Why study the Philippine biodiversity?
- **Why analyze the calls/signals?**

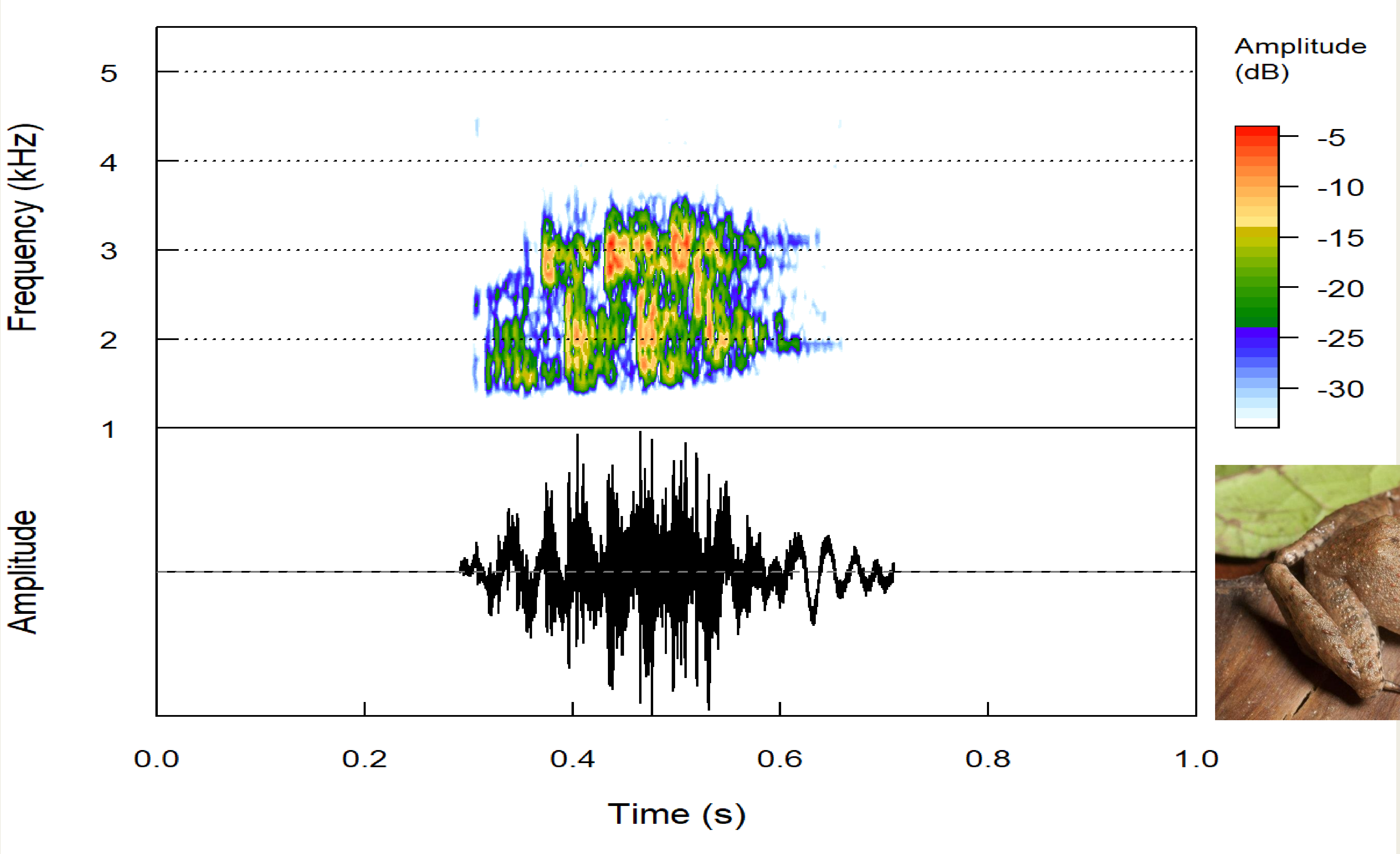
  - **Sexual selection:** Sexual selection acts on an organism's ability to obtain (often by any means necessary!) or copulate successfully with a mate
  - **Advertisement calls**



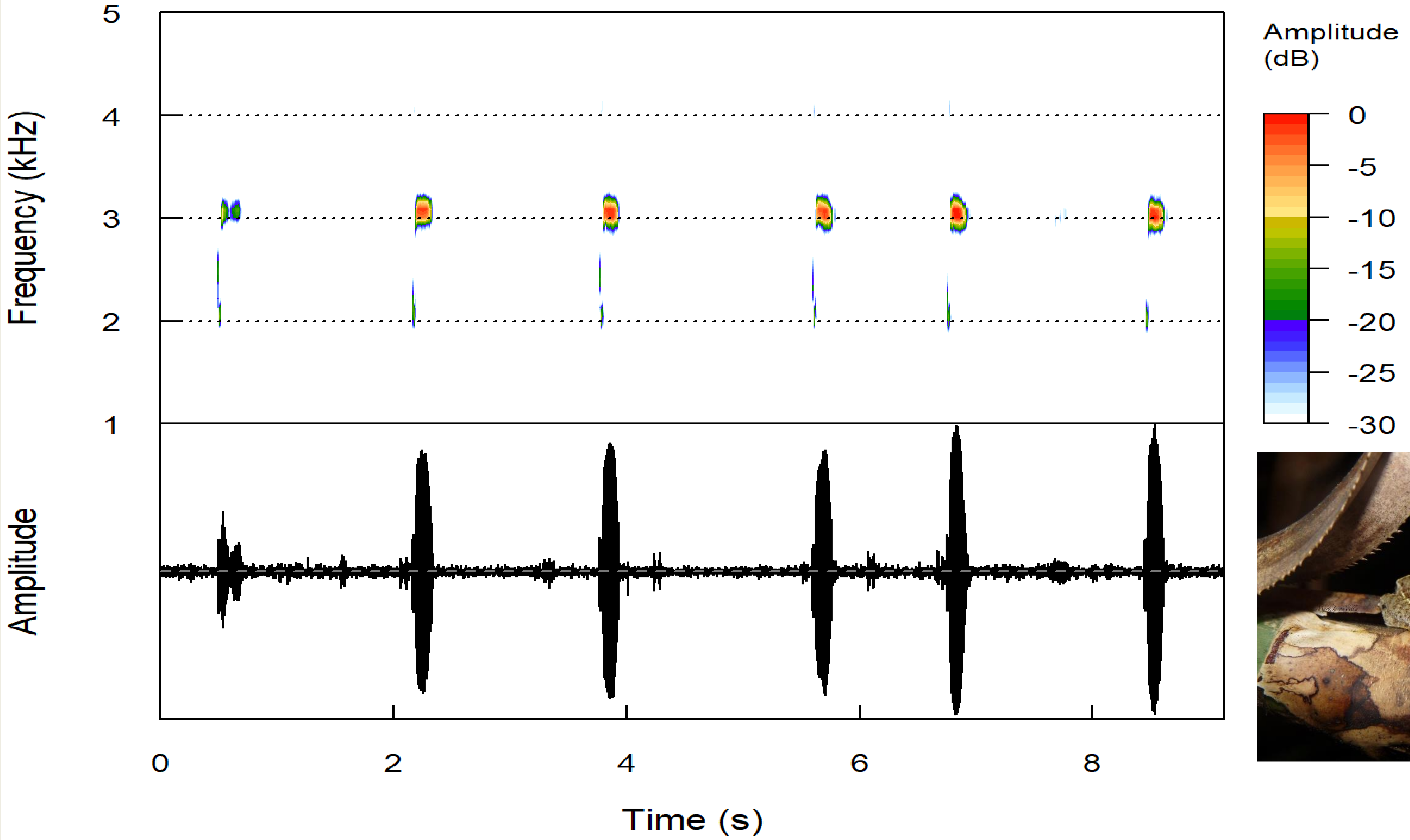
Platymantis cagayanensis (bout)



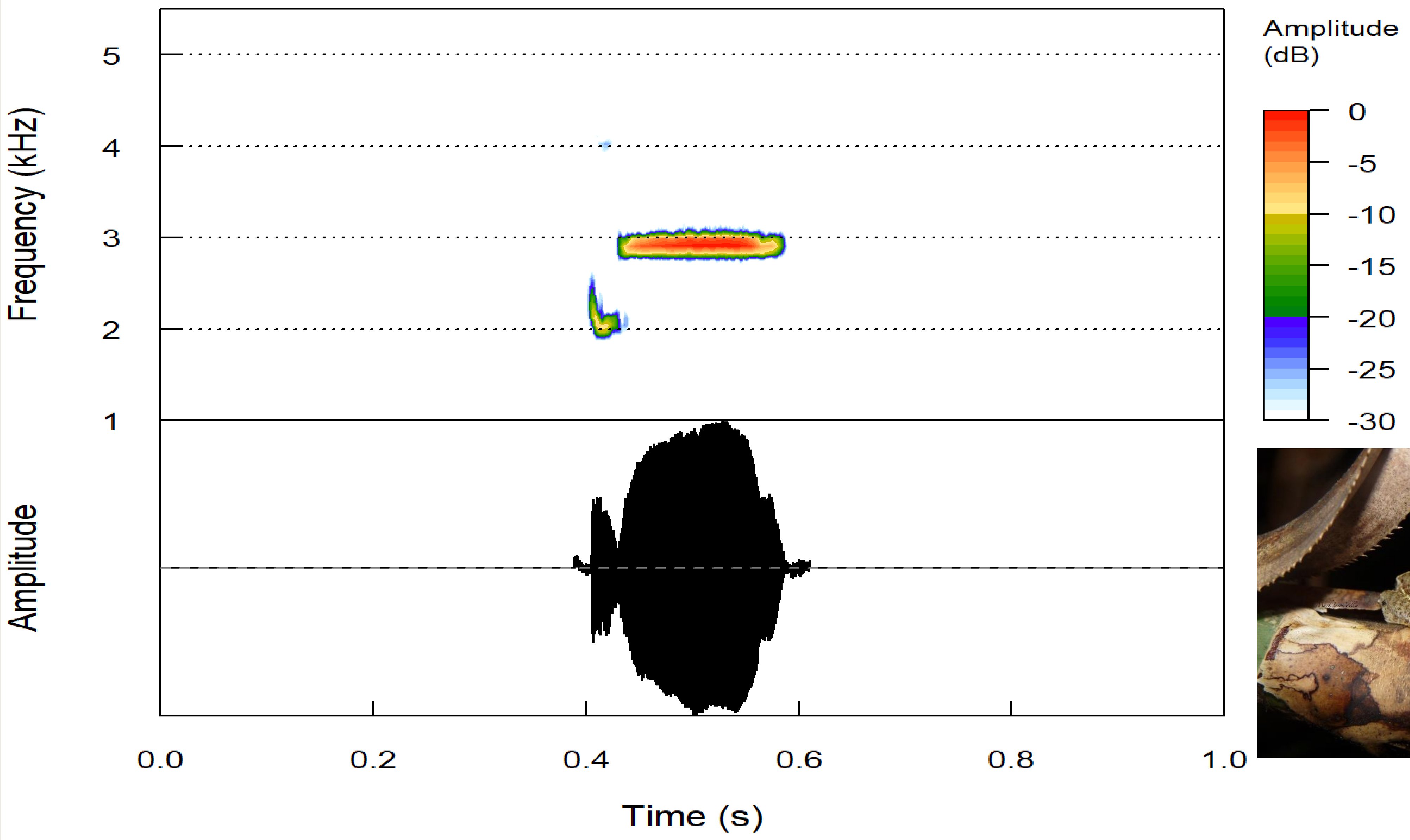
Platymantis cagayanensis (call/single note)



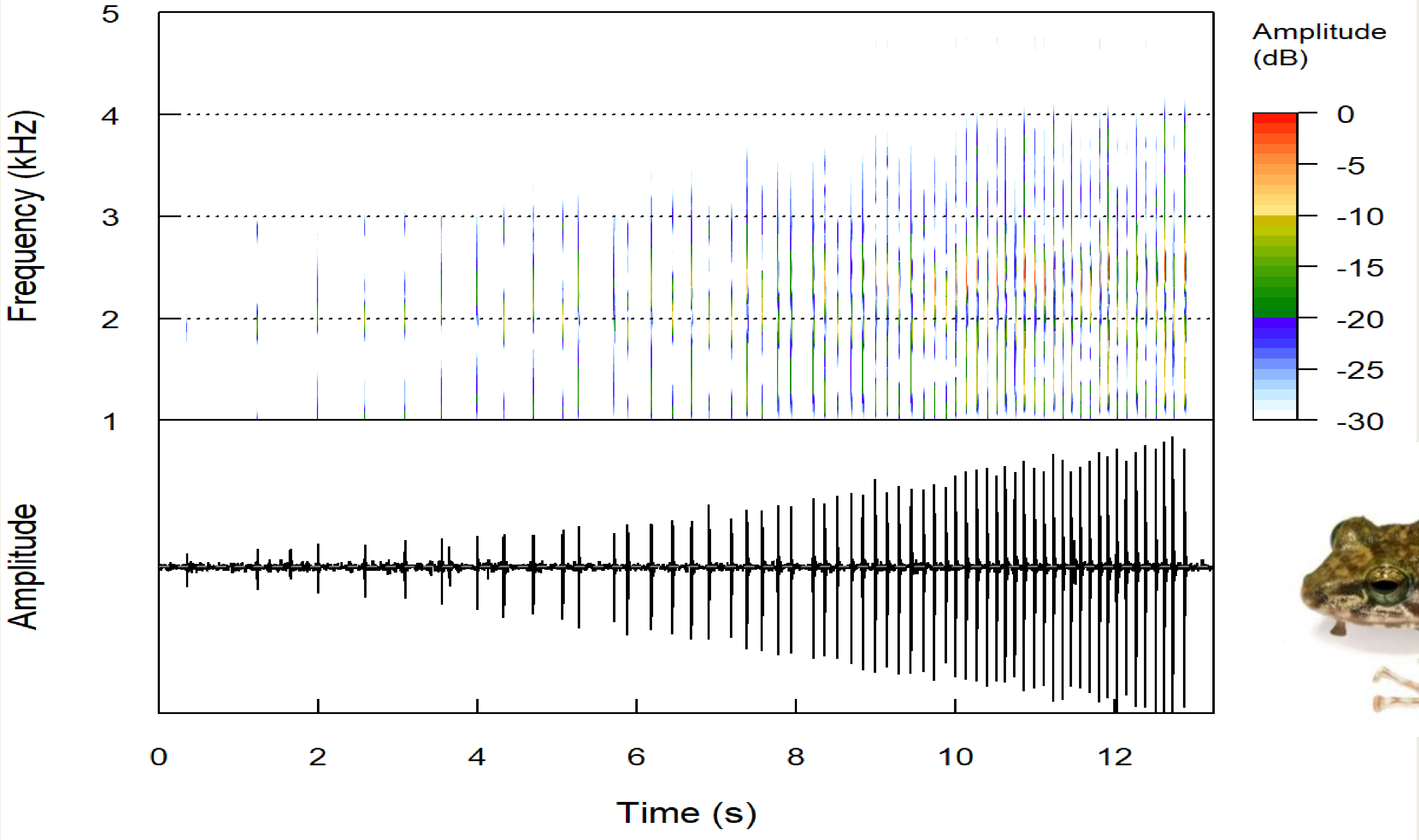
Platymantis isarog (random bout)



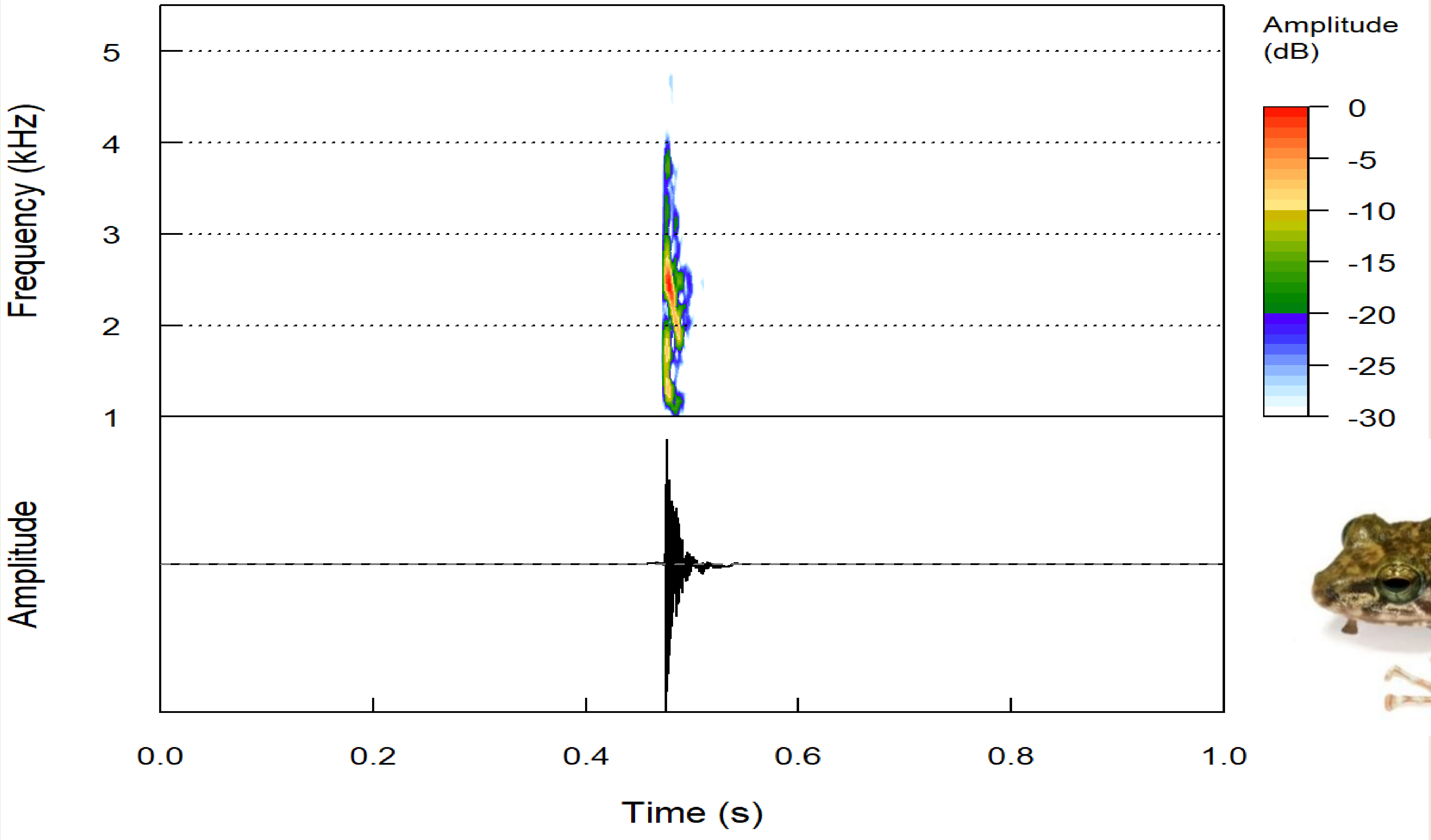
Platymantis isarog (call/single note)



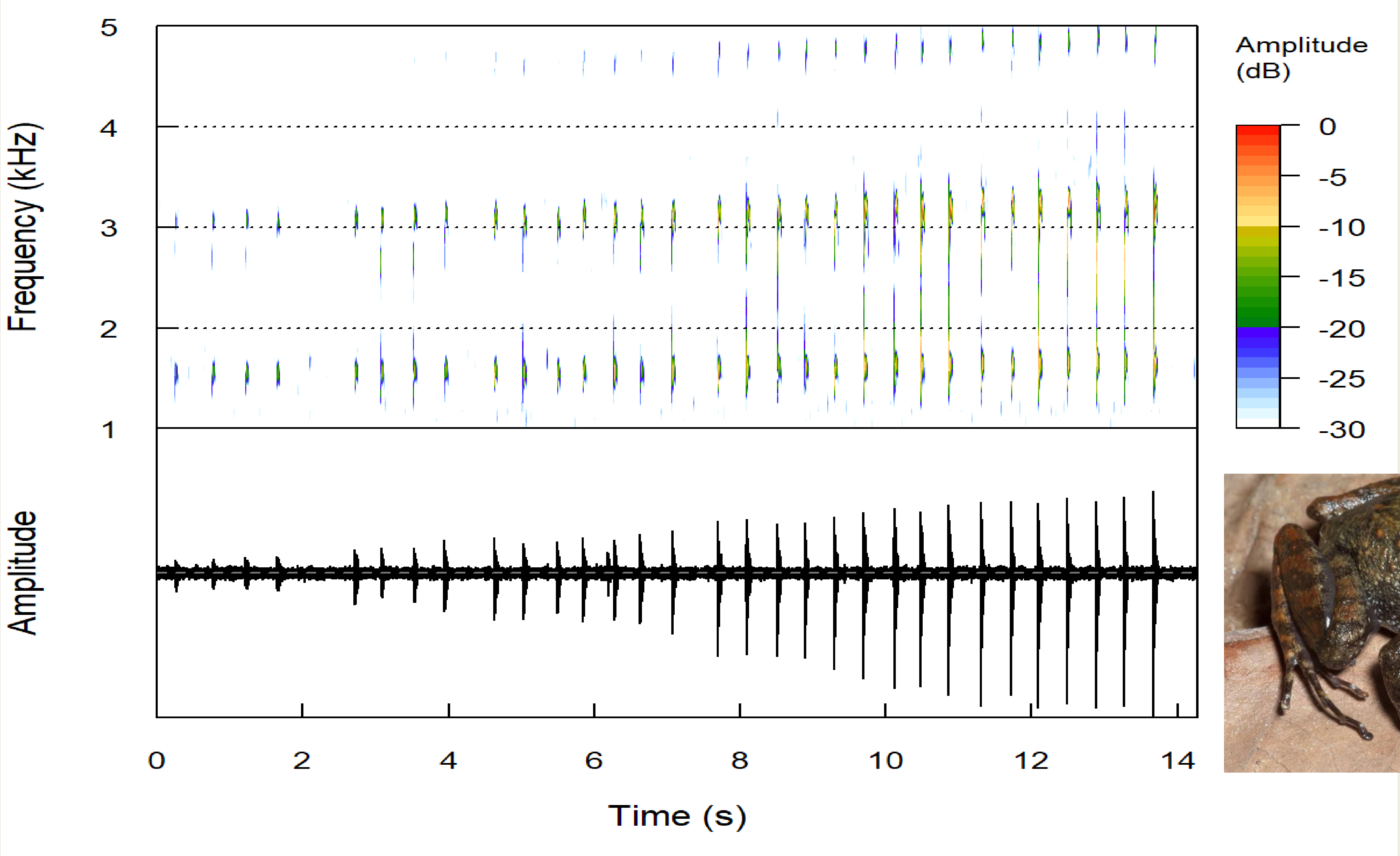
Platymantis insulatus (bout/call)



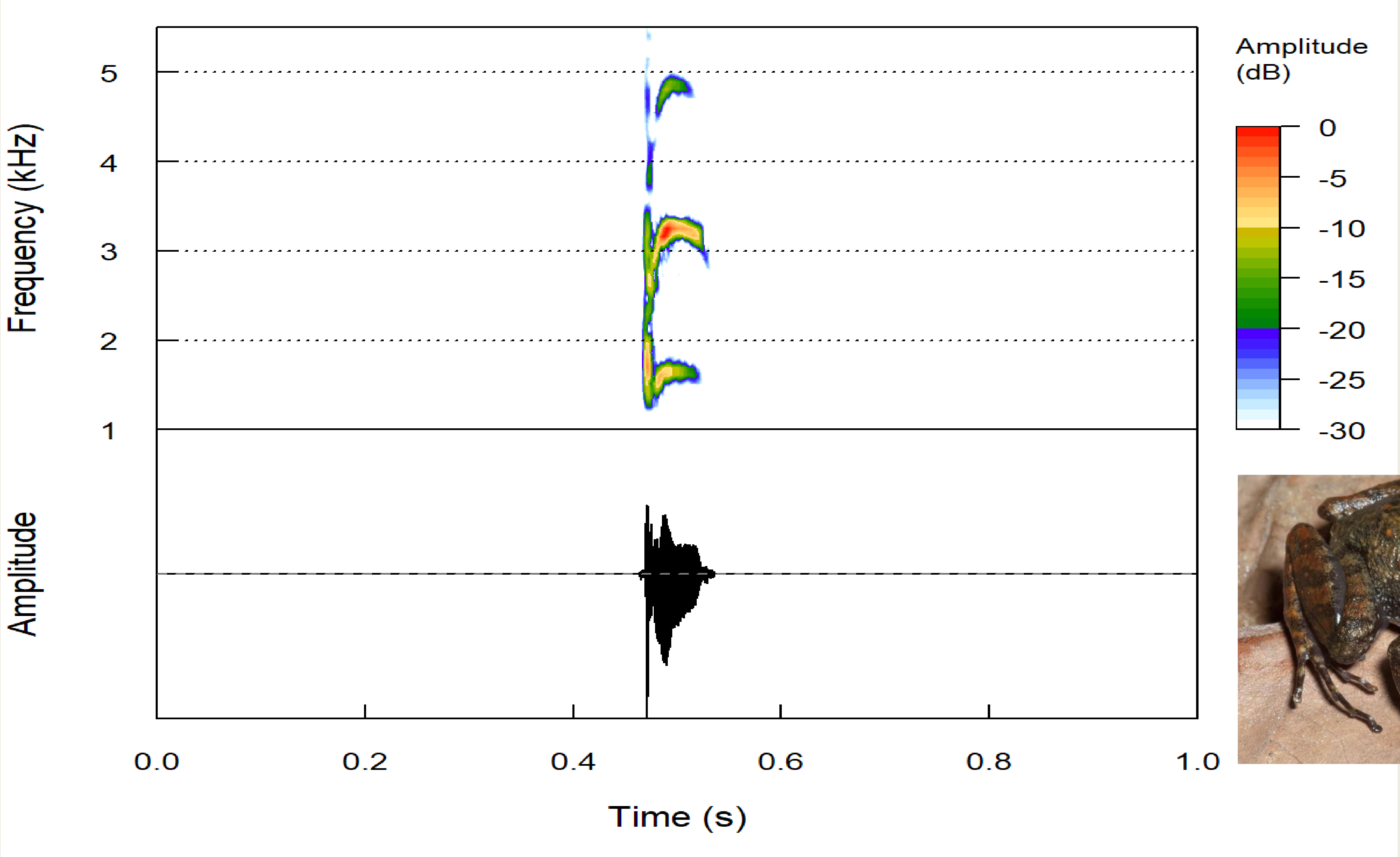
Platymantis insulatus (single note)



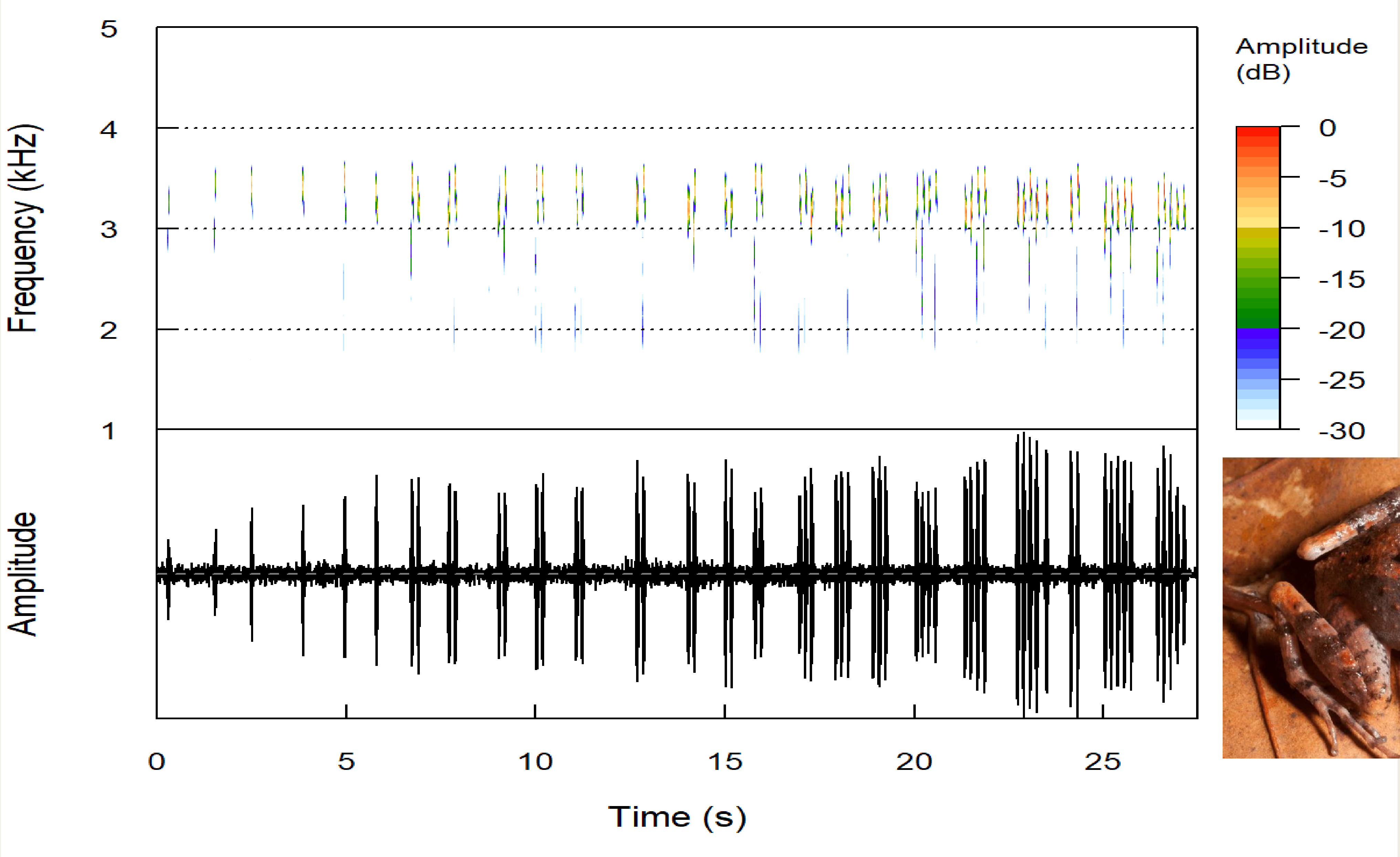
Platymantis levigatus (bout/call)



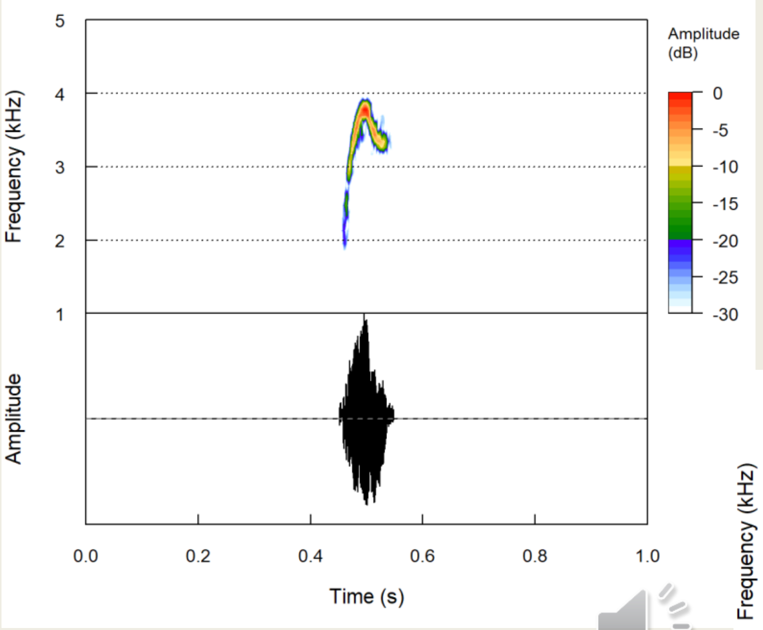
Platymantis levigatus (single note)

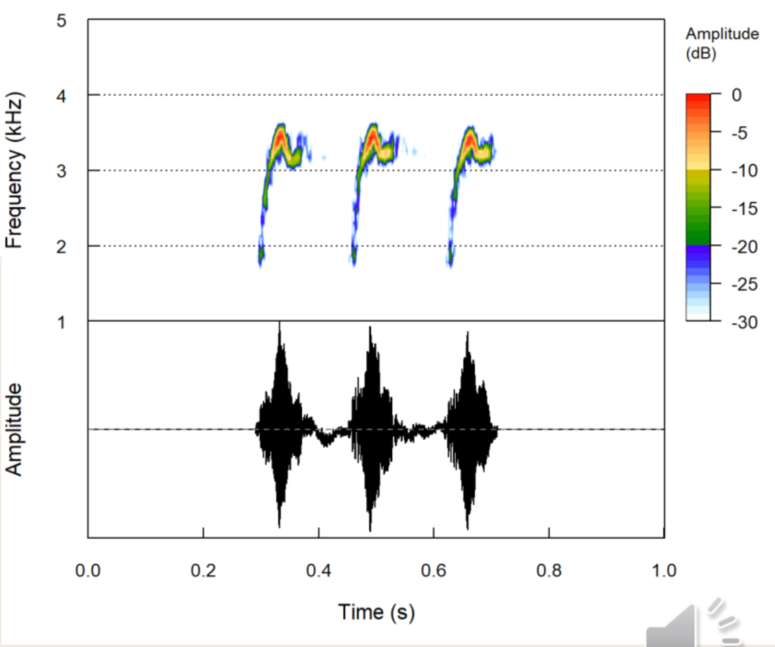


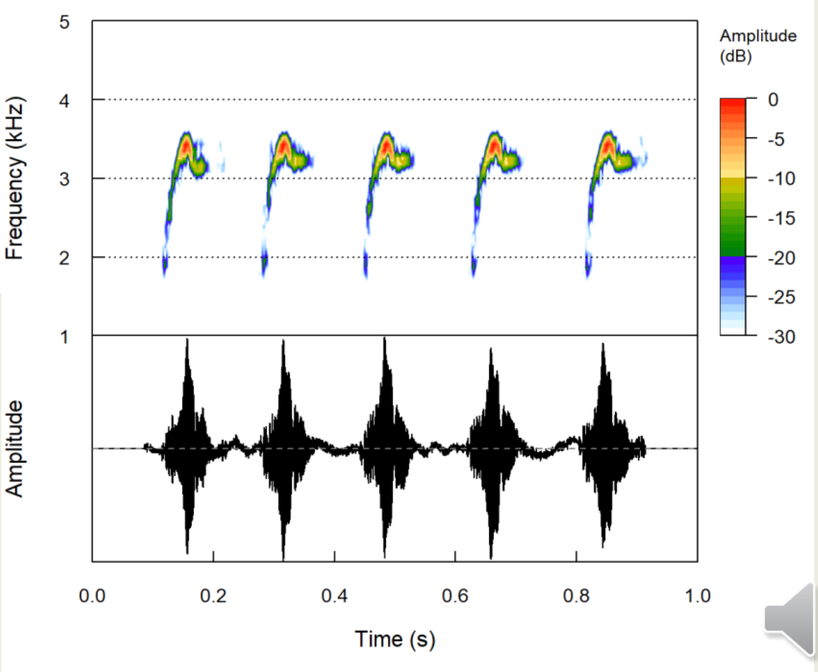
Shek shek (bout)



Shek shek (call)

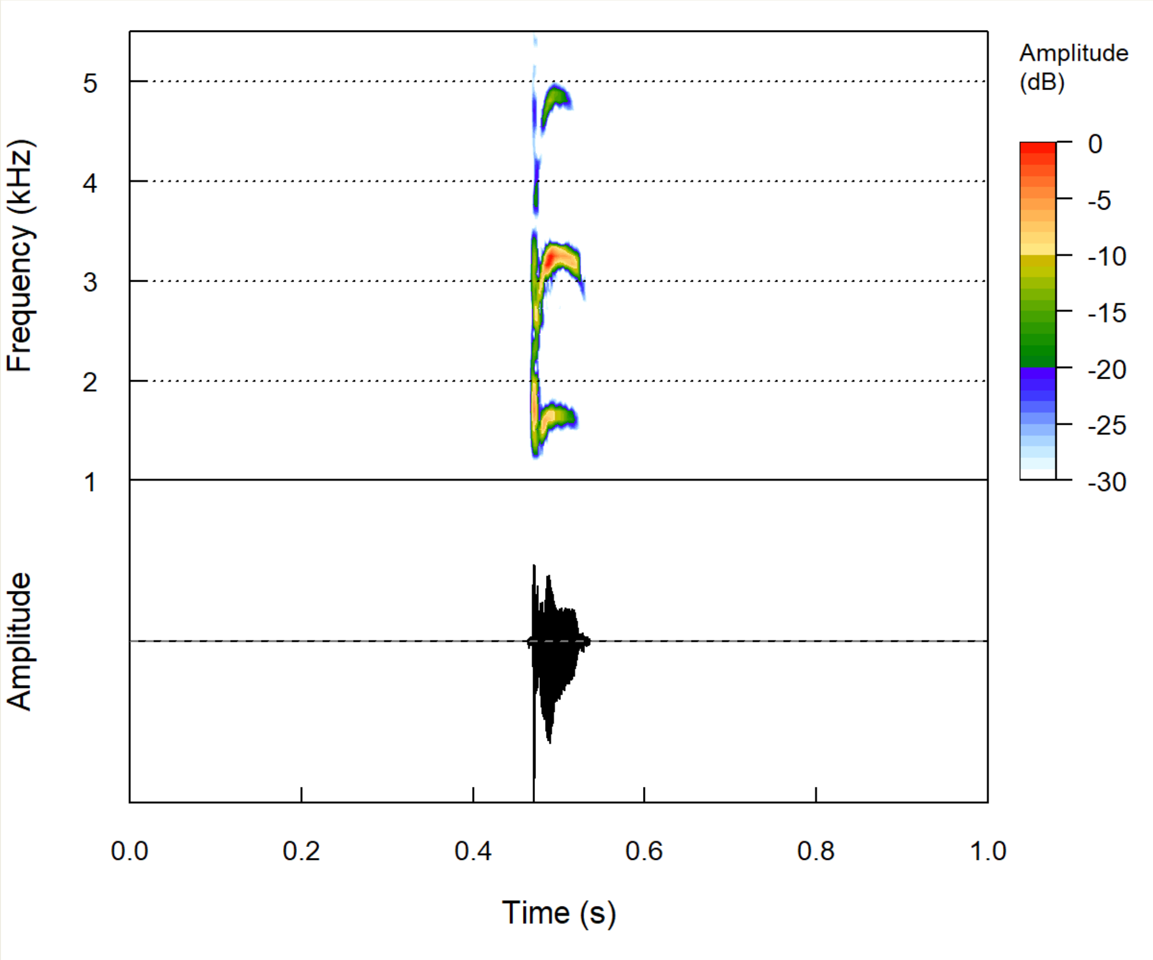






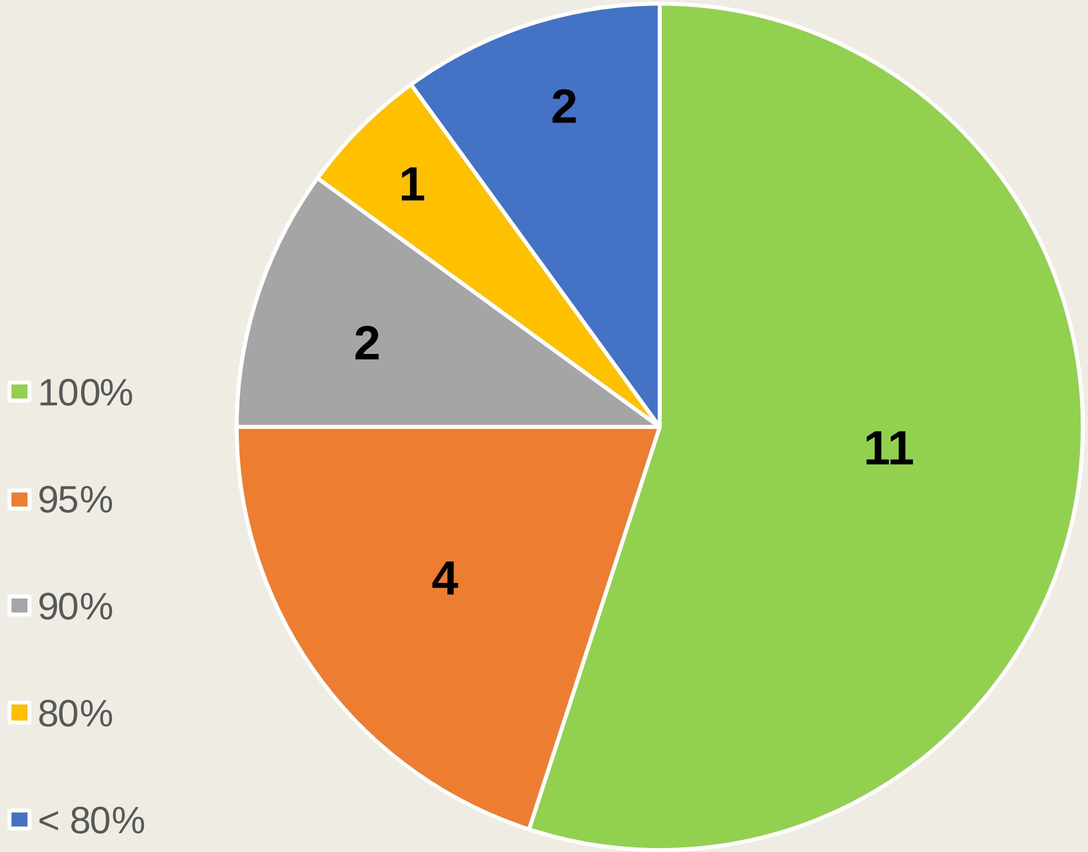
Methods

- Clipped single notes (20 samples per sp.)
  - Added silence to have the same time length
  - Same frequency range to cover all species
- Trained TensorFlow on 20 described species
- 20 described sp. Vs 21 undescribed sp.
- Island-based identification



Results

20 described species → Overall Rate: 94.3%

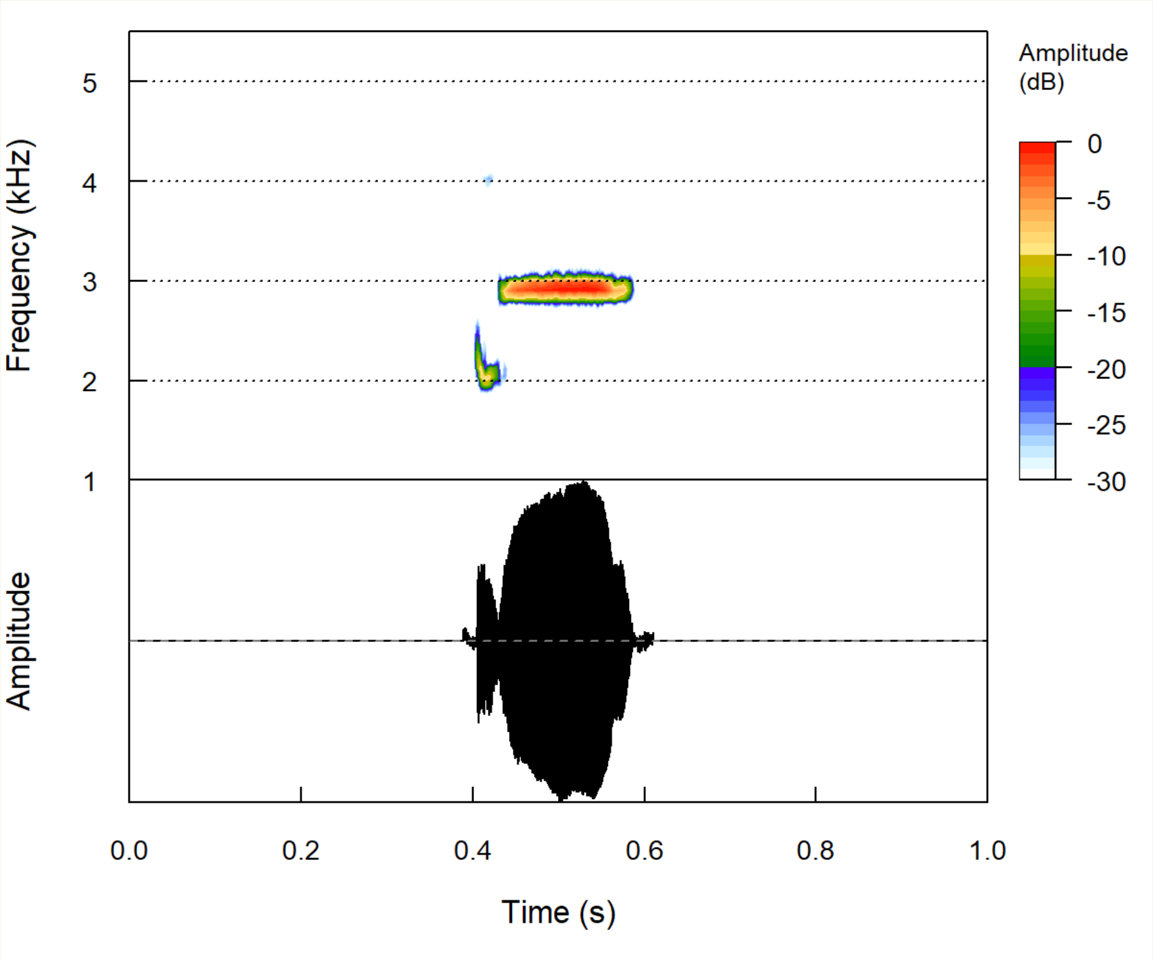


Described vs undescribed species

- TF identified all 20 *P. isarog* calls as isarog
  - TF certainty = 98.1%

- TF identified all 20 *P. diesmosi* calls as *P. insulates*
  - TF certainty = 94.3%

- TF confused "churink redori" with 10 different species
  - TF certainty = 38.4%

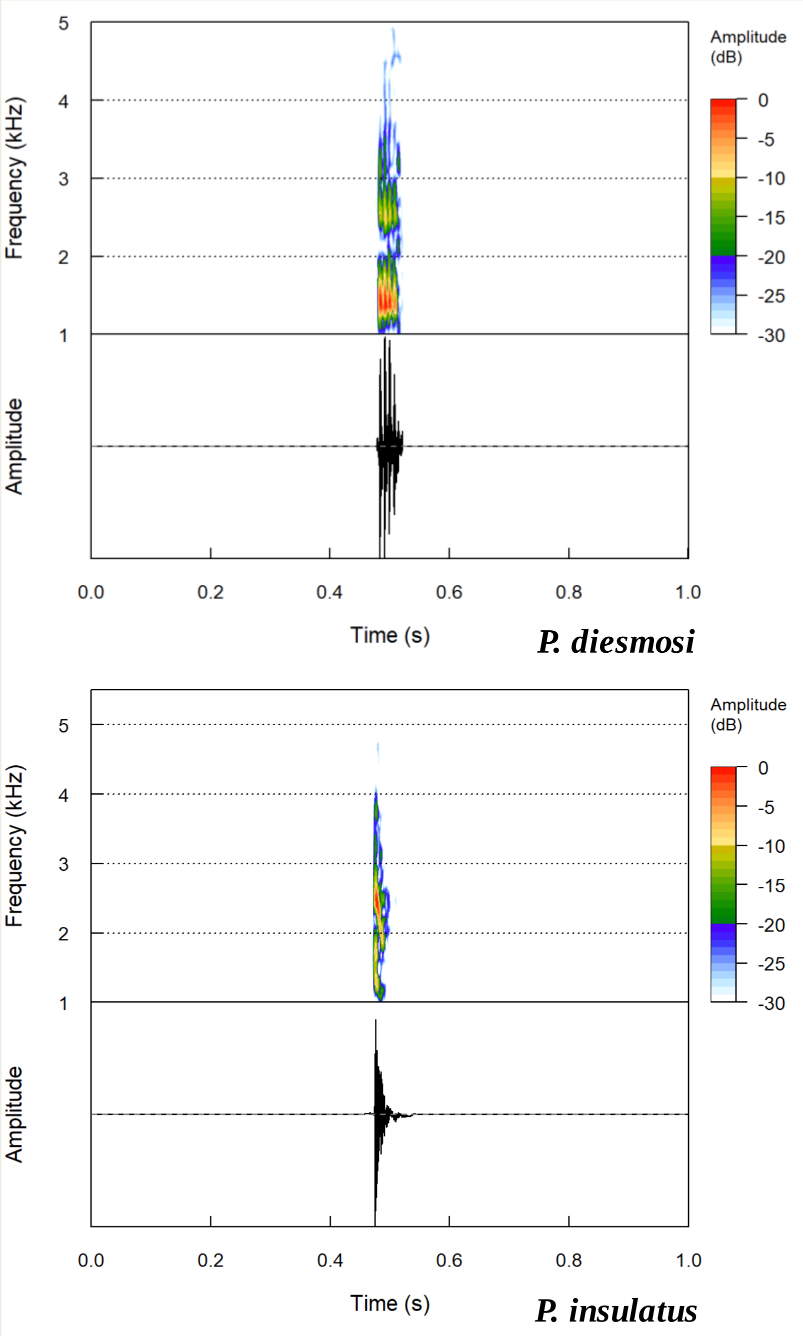


Described vs undescribed species

- TF identified all 20 *P. isarog* calls as isarog
  - TF certainty = 98.1%

- **TF identified all 20 *P. diesmosi* calls as *P. insulates***
  - **TF certainty = 94.3%**

- TF confused "churink redori" with 10 different species
  - TF certainty = 38.4%

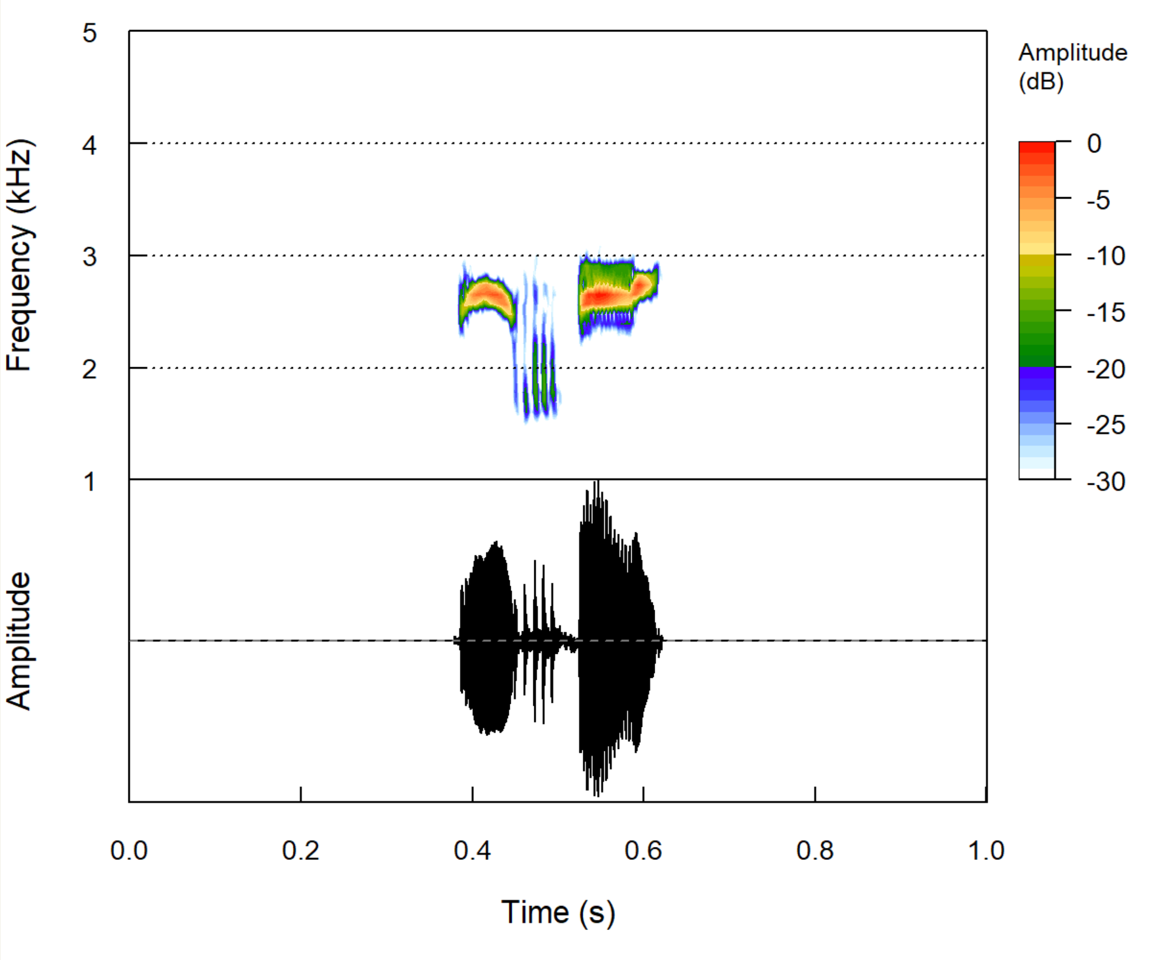


Described vs undescribed species

- TF identified all 20 *P. isarog* calls as isarog
  - TF certainty = 98.1%

- TF identified all 20 *P. diesmosi* calls as *P. insulates*
  - TF certainty = 94.3%

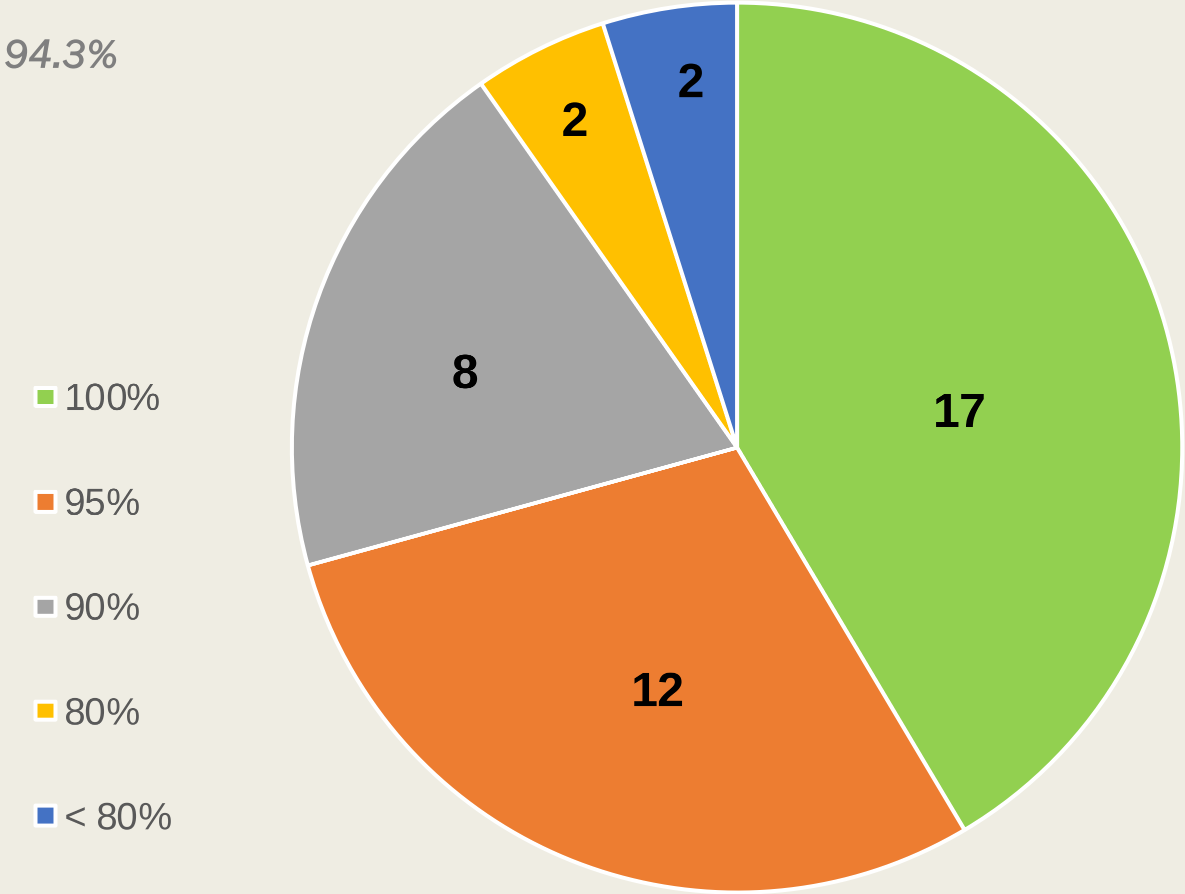
- **TF confused "churink redori" with 10 different species**
  - **TF certainty = 38.4%**



Results

All 41 species → Overall Rate: 94.1%

20 species → Overall Rate: 94.3%



Island-based Identification

Island names	Number of species	Overall ID rate
Luzon	27	94.60%
Mindanao	6	100%
Sibuyan | Tablas	6	100%
Polillo	6	99.20%
Samar | Leyte | Bohol	5	98.00%
Catanduanes	4	100%
Negros	4	100%
Dinagat | Siargao	4	96.30%
Gigante	3	100%
Panay	3	100%
Romblon	3	100%

The overall average ID rate across all islands = 98.7%

Outline

- Introduction
- Morphology-based Identification
  - **Project 1: Automated identification of Chagas disease vectors using statistical classifiers** (Completed Project)
  - **Project 2: TensorFlow improves automated identification of Chagas disease vectors** (Completed Project)
  - **Project 3: Marshalling diverse big data streams to understand risk of tick-borne diseases in the Great Plains** (Future Project)
- Signal-based Identification
  - **Project 1: Adapting TensorFlow to improve biodiversity assessment for Philippine frog species** (Current Project)
  - **Project 2: TensorFlow helps surveillance of mosquito species using cell phone recordings of wingbeats** (Current Project)
- Conclusion

Mosquito-borne diseases

- ❖ **Malaria**
- ❖ **Zika virus**
- ❖ **West Nile virus**
- ❖ **Chikungunya virus**
- ❖ **Dengue fever**




Mosquito wingbeats

❖ Helps to identify species

❖ Gives information about the sex

❖ Helps surveillance of mosquito species

❖ Provides the opportunity to control mosquito populations

Methods

~300 recordings over the past mosquito seasons

40 samples of *Aedes aegypti* from Mexico

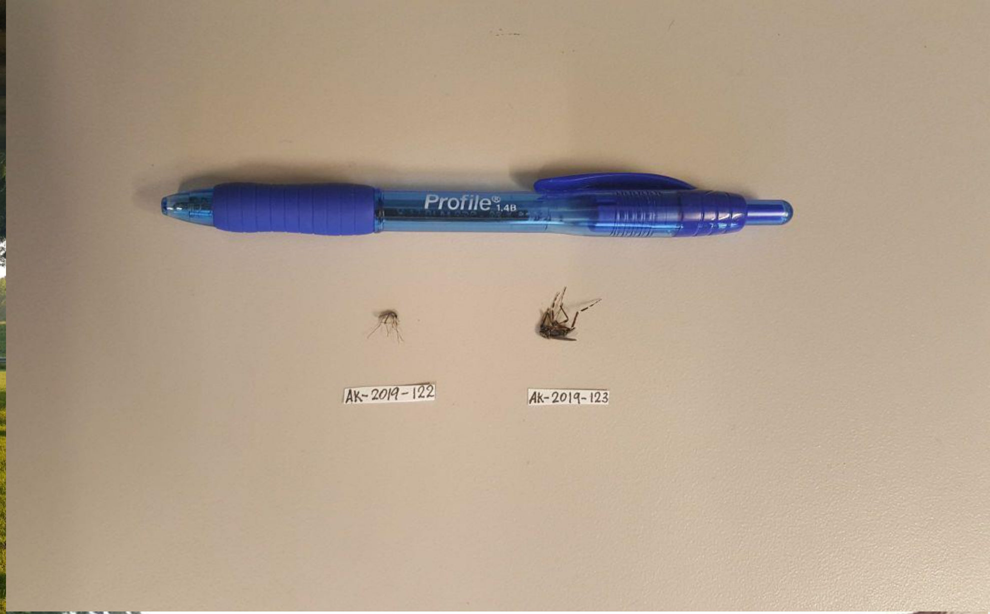
16 samples of *Anopheles gambiae* from Ghana

Collecting instruction videos in English and Spanish

Goals

1) Identifying species in Kansas mosquito community

2) Detecting the presence of two alien species



Outline

- Introduction
- Morphology-based Identification
  - **Project 1: Automated identification of Chagas disease vectors using statistical classifiers** (Completed Project)
  - **Project 2: TensorFlow improves automated identification of Chagas disease vectors** (Completed Project)
  - **Project 3: Marshalling diverse big data streams to understand risk of tick-borne diseases in the Great Plains** (Future Project)
- Signal-based Identification
  - **Project 1: Adapting TensorFlow to improve biodiversity assessment for Philippine frog species** (Current Project)
  - **Project 2: TensorFlow helps surveillance of mosquito species using cell phone recordings of wingbeats** (Current Project)
- Conclusion

❖ **Achievements**

- *T. dimidiata* HG1, HG2, HG3
- *P. dorsalis* vs *P. guntheri*
- Citizen-scientists

HG1 HG2 HG3

➢ **Challenges**

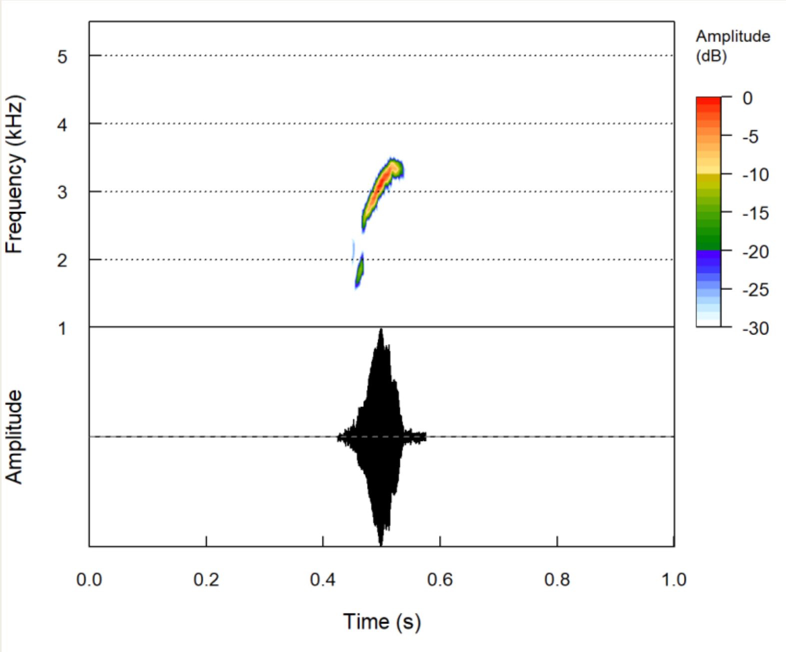
➢ **Caveats**

❖ Achievements

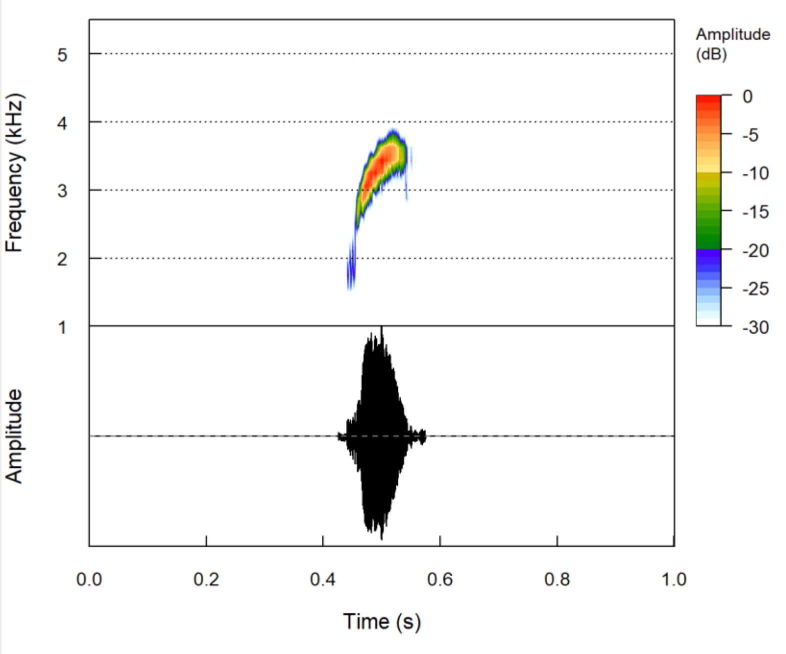
- *T. dimidiata* HG1, HG2, HG3
- ***P. dorsalis* vs *P. guntheri***
- Citizen-scientists

➢ Challenges

➢ Caveats



P. dorsalis



P. guntheri

- **Achievements**
  - ***T. dimidiata* HG1, HG2, HG3**
  - ***P. dorsalis* vs *P. guntheri***
  - **Citizen-scientists**

- **Challenges**
- **Caveats**

➢ Achievements

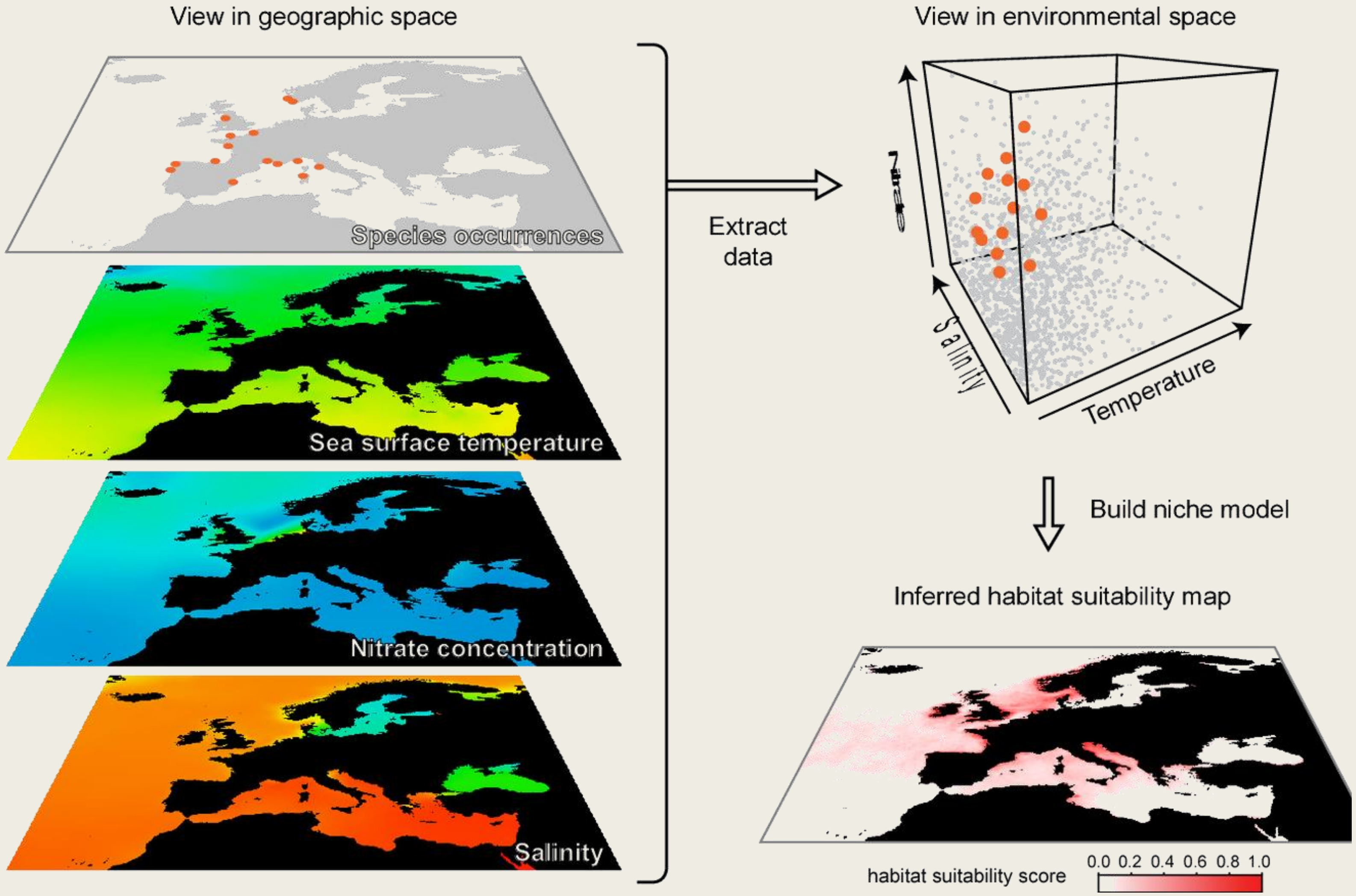
❖ **Challenges**

- **Correct identification of species**
- **Availability of data**

➢ Caveats

- ➢ Achievements
- ➢ Challenges
- ❖ **Caveats**

  - **Number of images**
  - **Concepts vs tools**



Acknowledgement

❖ **Co-authors**

- Town Peterson
- Ed Komp
- Janine Ramsey
- Rodrigo Gurgel-Gonçalves,
- Rafe Brown
- Johana Goyes
- Nathan Burkett-Cadena

❖ **Collaborators**

- Daniel Jiménez-García
- Lindsay Campbell
- Spencer Mattingly