MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

Anuran species recognition using a hierarchical classification approach

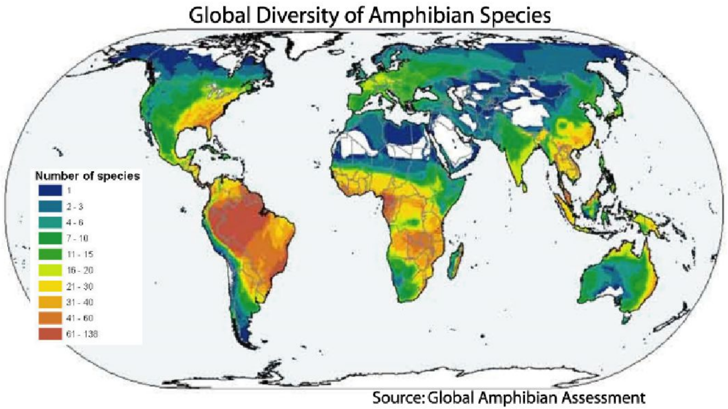
Juan G. Colonna[12], João Gama[2], and Eduardo F. Nakamura[1]

[1]Federal University of Amazonas (UFAM), Institute of Computing (Icomp)
[2]Laboratory of Artificial Intelligence and Decision Support (LIAAD), INESC Tec
{juancolonna, nakamura}@icomp.ufam.edu.br
jgama@fep.up.pt

Getting more from family, genus and species of frogs

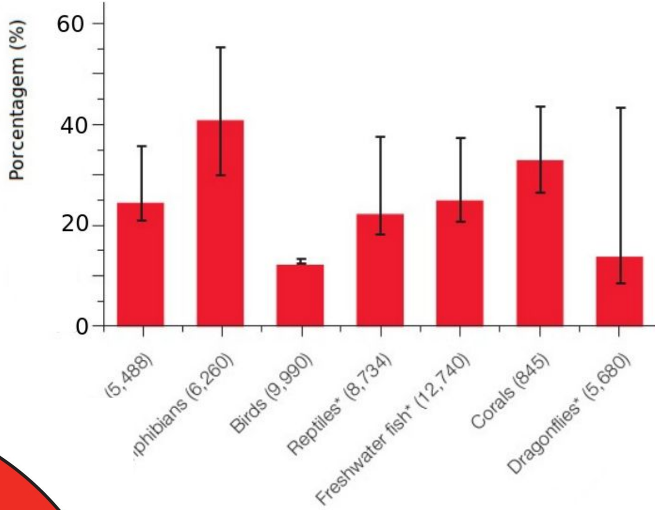
Introduction - Why frogs?

- **Anura** is the name of an order of animals in the **Amphibian** class which lack a tail, this includes **frogs** and **toads**.

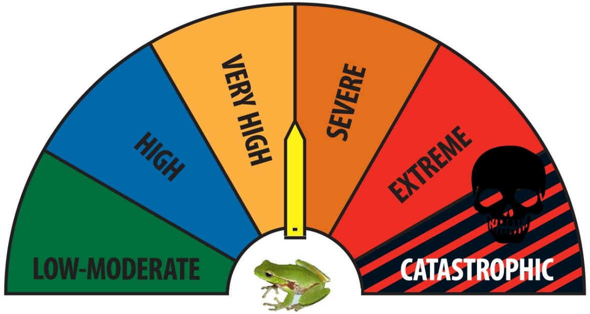


Global Diversity of Amphibian Species

Source: Global Amphibian Assessment



- **Frogs are very sensitive to environmental changes**



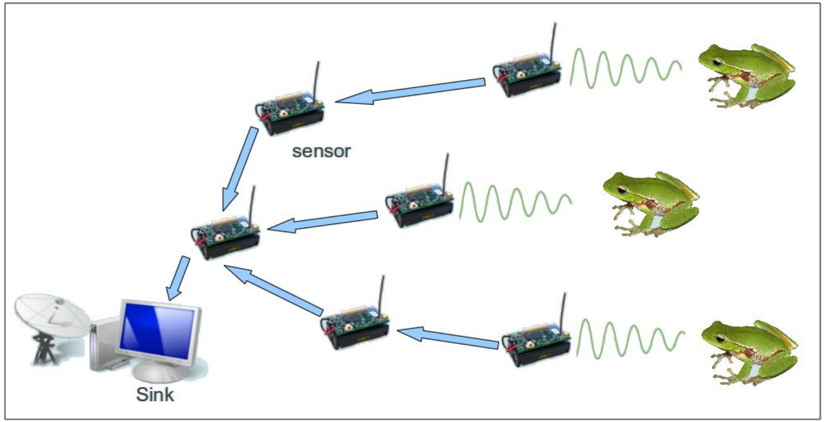
Why monitor populations of frogs?

Hypothesis: Tracking the changes in the anuran populations can help us to determine ecological problems in early stages.

It involves several manual tasks!

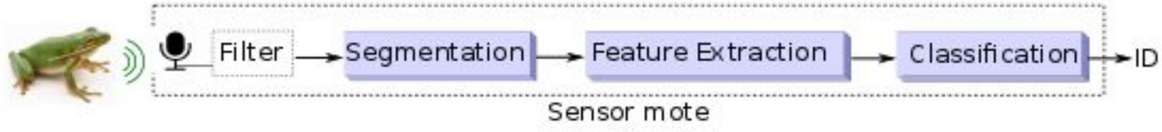
Proposal

Signal processing (SP) + Wireless Sensor Networks (WSN) + Machine Learning (ML)



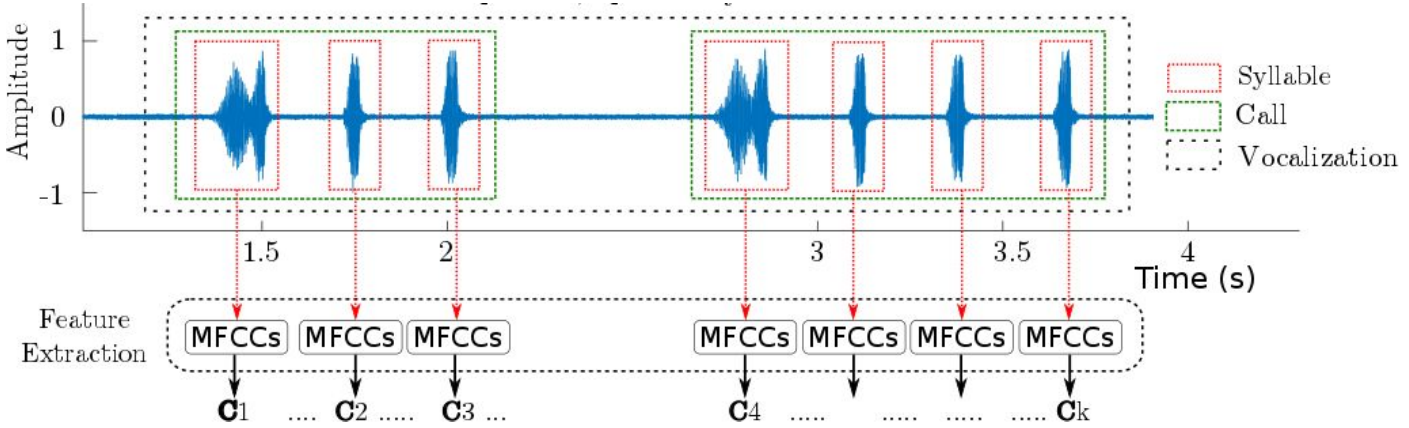
Advantages: It is Automatic, less intrusive and allows long term monitoring.

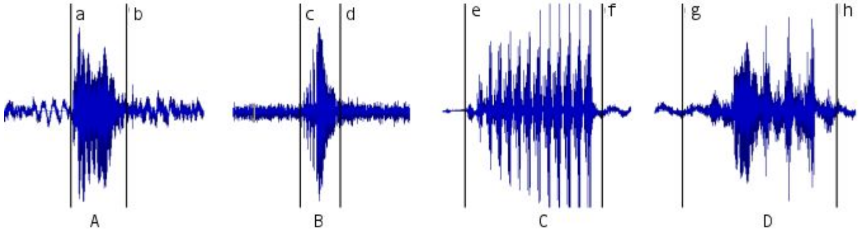
How to do that?

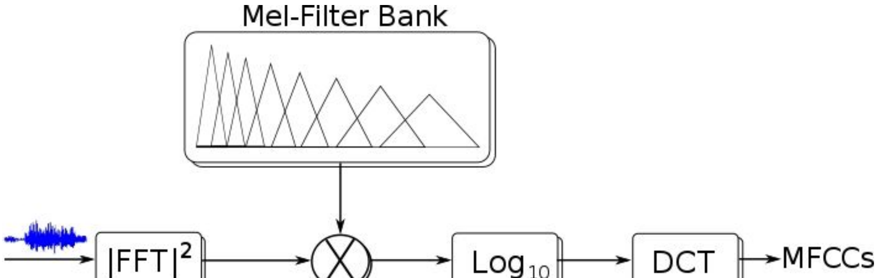


1) Pre-processing:
    a) Filter: band-pass filter, wavelet decomposition, etc.
    b) Segmentation: **syllable-based approach** (x_k)
2) Feature Extraction: that maps $x_k \rightarrow c_k$
    a) Mel-frequency cepstral coefficients (MFCCs)
    b) Spectral centroid, Spectral bandwidth, Pitch, etc.
3) Recognition: ML technique to classify $c_k \rightarrow ID$ (species ID)
    a) Support Vector Machine, kNN, Tree, etc.

Segmentation and feature extraction





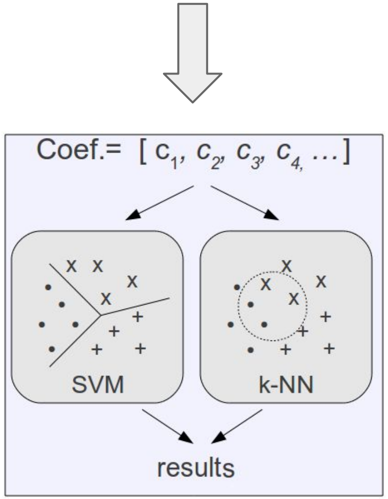


Traditional Classification Approach

Dataset with:

- *k* samples (or syllables)
- *l* coefficients
- one label (s_j={j species})

$$\text{dataset} = \begin{bmatrix} \mathbf{c}_1 = [c_1, c_2, \dots, c_l], & s \\ \mathbf{c}_2 = [c_1, c_2, \dots, c_l], & s \\ \vdots & \vdots \\ \mathbf{c}_k = [c_1, c_2, \dots, c_l], & s_j \end{bmatrix}$$

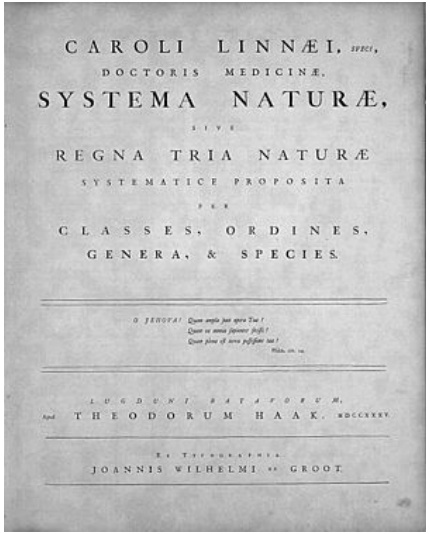


Then, apply a **“flat”** classifier (kNN, SVM, etc.)

Problem: the number of classes grows together with the number of species who wish to recognize increasing the complexity of the model.

Knowledge organization

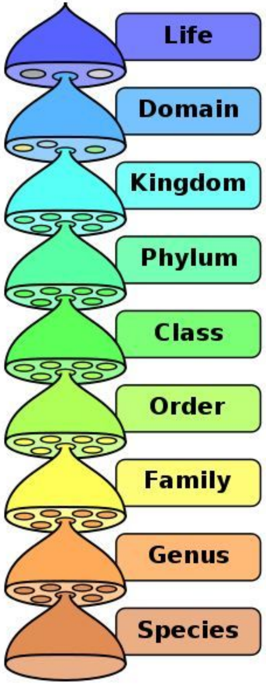
Carl Linnaeus has defined a particular form of biological organization called *taxonomy* in his work *Systema Naturae* (1735).



CAROLI LINNÆI, SVECI,
DOCTORIS MEDICINÆ,
SYSTEMA NATURÆ,
SIVE
REGNA TRIA NATURÆ
SYSTEMATICE PROPOSITA
PER
CLASSES, ORDINES,
GENERA, & SPECIES.

LUGDUNI BATAVORUM,
Apud THEODORUM HAAK. MDCCXXXV.

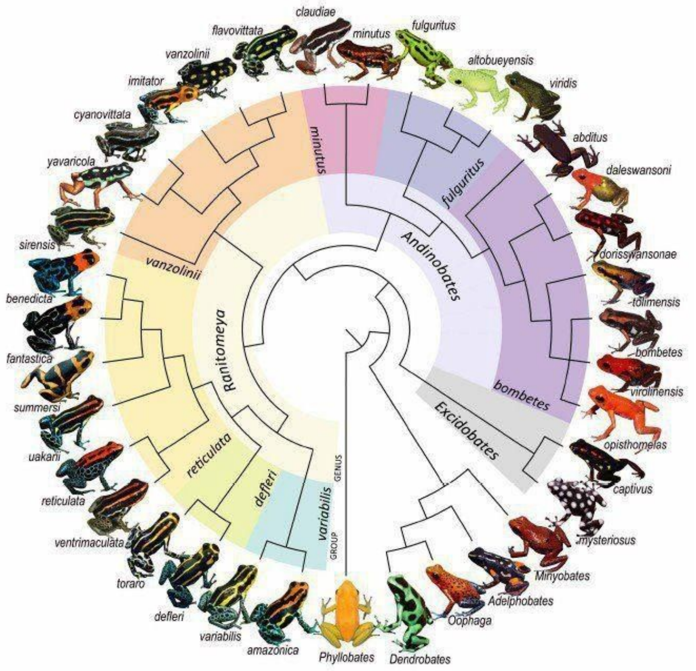
Ex Typographia
JOANNIS WILHELMI de GROOT.



How to improve the classification using the taxonomy?

- The anura Order has 31 Families (approximately)
- These Families are divided into several genus
- And finally, these genus are divided in almost 6000 species

Hypothesis: the phylogenetic taxonomy may describe similar calls among species that belong to the same genus and family[2].



Illustrative figure.

---

[2] B. Gingras and W. T. Fitch.
A three-parameter model for classifying anurans into four genera based on advertisement calls.
The Journal of the Acoustical Society of America, 133(1):547–559, 2013.

A Multi-output approach (multi-class and multi-label)

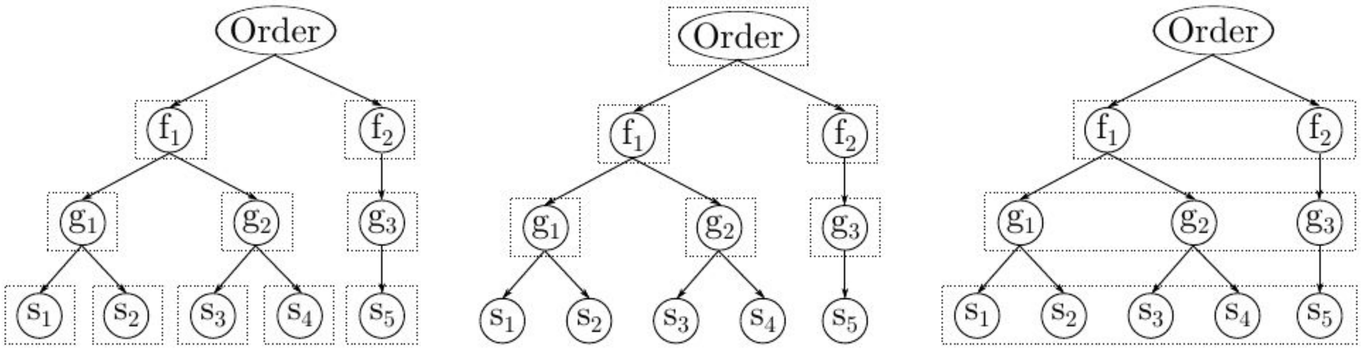
Extend the dataset incorporating the new labels:

- Label s_j = {j different species}
- Label g_i = {i different genus}
- Label f_m = {m different families}

$$\text{Multi-label dataset} = \begin{bmatrix} \mathbf{c}_1 = [c_1, c_2, \ldots, c_l], & s, & g, & f \\ \mathbf{c}_2 = [c_1, c_2, \ldots, c_l], & s, & g, & f \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{c}_k = [c_1, c_2, \ldots, c_l], & s_j, & g_i, & f_m \end{bmatrix}$$

Hierarchical problem decomposition

- Use the taxonomy relation of the labels to build a tree.



(a)

One classifier per node

(b)

One classifier per parent node

(c)

One classifier per level

Our dataset

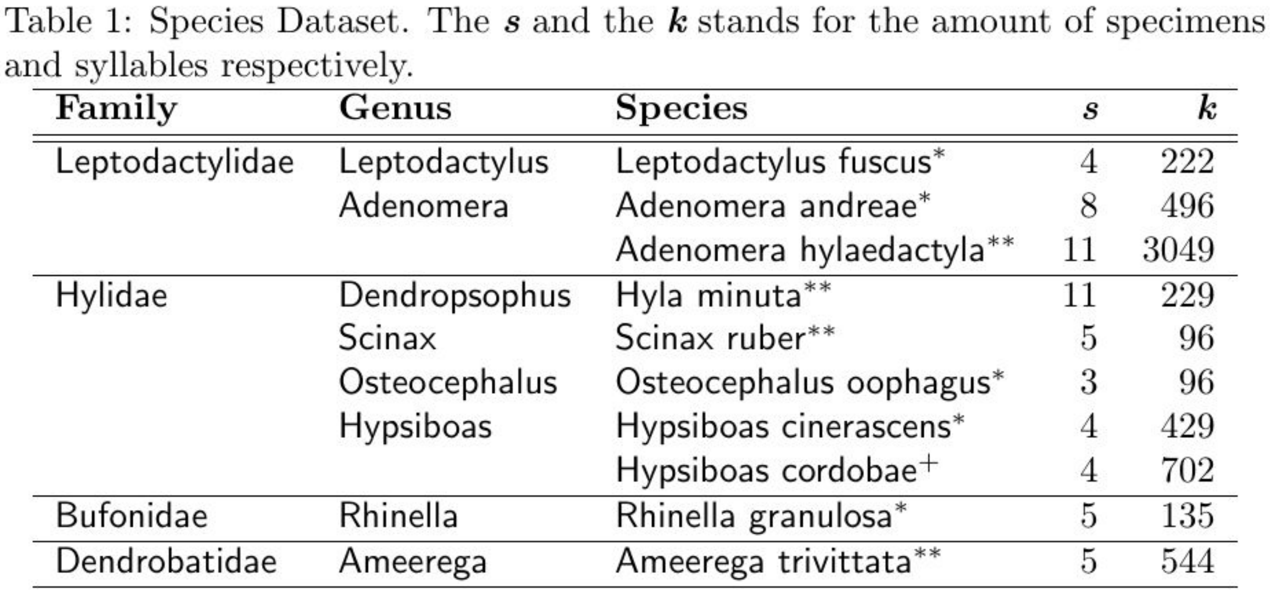
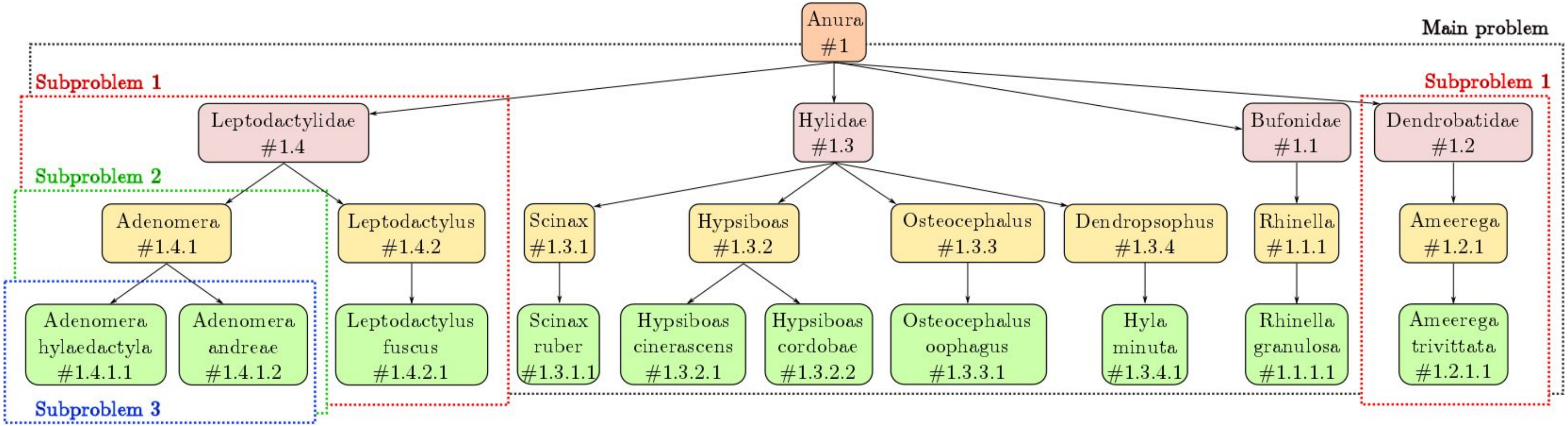


Table 1: Species Dataset. The s and the k stands for the amount of specimens and syllables respectively.

Family	**Genus**	**Species**	s	k
Leptodactylidae	Leptodactylus	Leptodactylus fuscus*	4	222
	Adenomera	Adenomera andreae*	8	496
		Adenomera hylaedactyla**	11	3049
Hylidae	Dendropsophus	Hyla minuta**	11	229
	Scinax	Scinax ruber**	5	96
	Osteocephalus	Osteocephalus oophagus*	3	96
	Hypsiboas	Hypsiboas cinerascens*	4	429
		Hypsiboas cordobae^{+}	4	702
Bufonidae	Rhinella	Rhinella granulosa*	5	135
Dendrobatidae	Ameerega	Ameerega trivittata**	5	544

- Indeed this is not a big-data dataset, but it is enough to prove our point.

Building our hierarchical classifier from our dataset

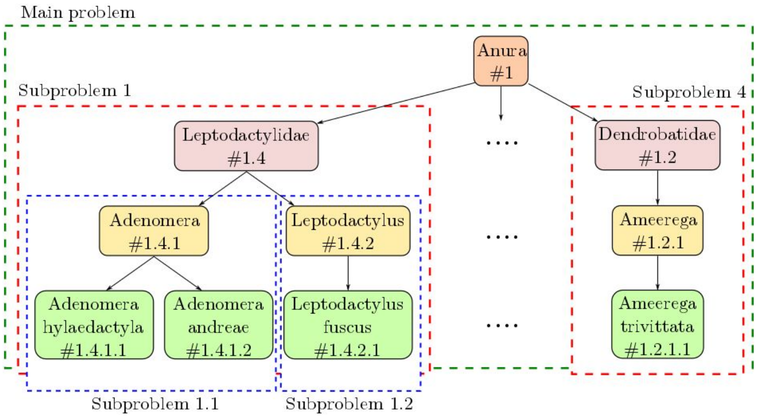


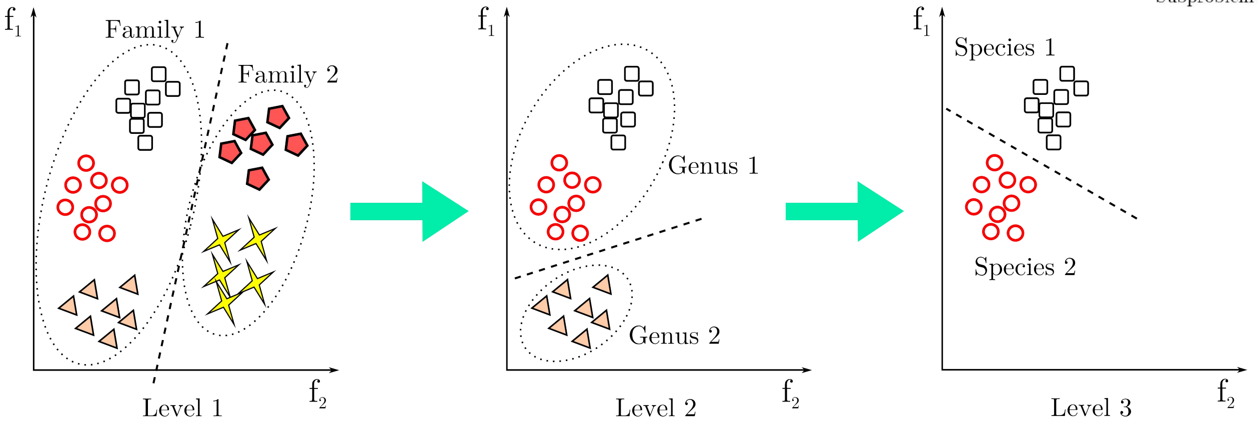
Benefit: One Classifier per Parent Node allows us to simplify the problem

Example: suppose that the first level decides in favor of the family *Bufonidae*. In this case there are no more splits in the tree, consequently it is not necessary to perform extra classifications to determine the species.

Hierarchical problem decomposition

- Subproblem decomposition and simplification:





Experiment configuration

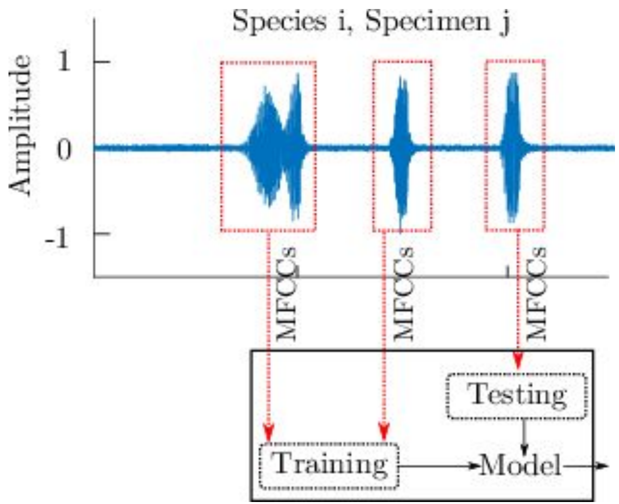
- A kNN was chosen as base classifier in each node (k=3).
- We adapted the cross-validation procedure to group syllables by individuals to test how well our method generalize.
- The Average-accuracy was used in evaluations to avoid an artificially increment of the Micro-accuracy due to unbalanced number samples in each class.

$$\text{Average-Acc} = \frac{1}{m}\sum_{i=1}^{m} \text{Acc}_i = \frac{1}{m}\sum_{i=1}^{m} \frac{tp_i}{k_i}$$

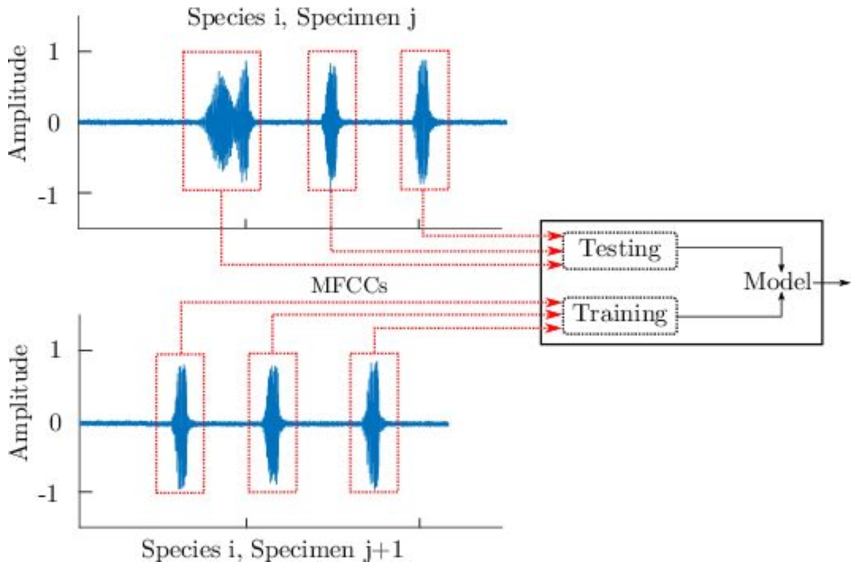
where Acc_i is the accuracy per row i of confusion matrix, m the total number of rows, tp_i are the true positives, and k_i the total number of syllables per row.

- **Random Baseline:**
  - Micro-accuracy = 0.50 (dummy classifier)
  - Average-accuracy = 0.10 (dummy classifier)

k-CV by Specimens (or individuals)



Common procedure found in the related works when syllable-based methodology is adopted.



Our Cross-Validation procedure by grouping syllables of the same individuals to test how well the model generalize.

Results per level

- Family Level (Acc = 76%)

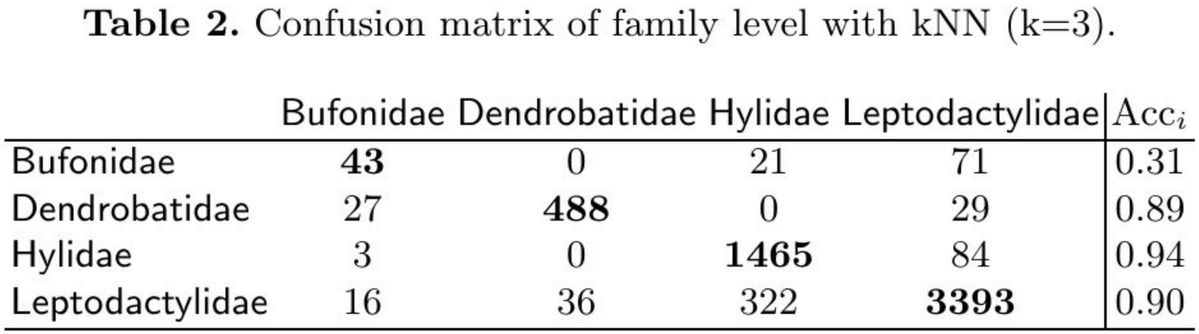


Table 2. Confusion matrix of family level with kNN (k=3).

	Bufonidae	Dendrobatidae	Hylidae	Leptodactylidae	Acc_i
Bufonidae	**43**	0	21	71	0.31
Dendrobatidae	27	**488**	0	29	0.89
Hylidae	3	0	**1465**	84	0.94
Leptodactylidae	16	36	322	**3393**	0.90

Results per level

- Family Level (Acc = 76%)
- Genus Level (Acc = 61%)

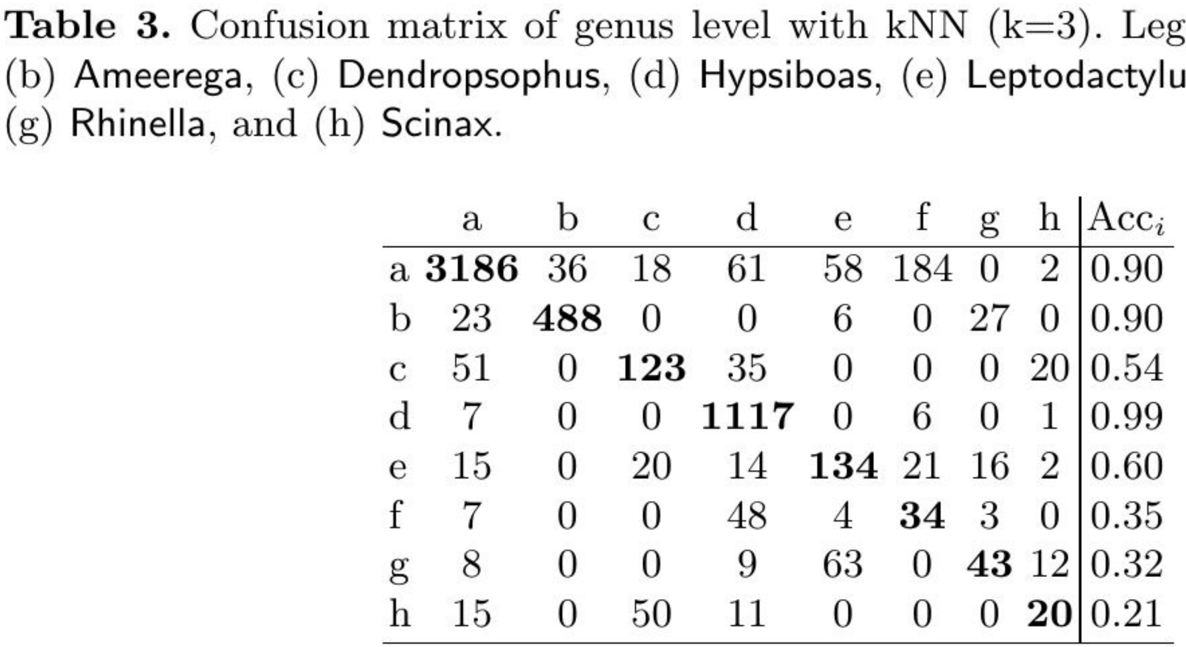


Table 3. Confusion matrix of genus level with kNN (k=3). Legend: (a) Adenomera, (b) Ameerega, (c) Dendropsophus, (d) Hypsiboas, (e) Leptodactylus, (f) Osteocephalus, (g) Rhinella, and (h) Scinax.

	a	b	c	d	e	f	g	h	Acc_i
a	**3186**	36	18	61	58	184	0	2	0.90
b	23	**488**	0	0	6	0	27	0	0.90
c	51	0	**123**	35	0	0	0	20	0.54
d	7	0	0	**1117**	0	6	0	1	0.99
e	15	0	20	14	**134**	21	16	2	0.60
f	7	0	0	48	4	**34**	3	0	0.35
g	8	0	0	9	63	0	**43**	12	0.32
h	15	0	50	11	0	0	0	**20**	0.21

Results per level

- Family Level (Acc = 76%)
- Genus Level (Acc = 61%)
- Species Level (Acc = 61%)

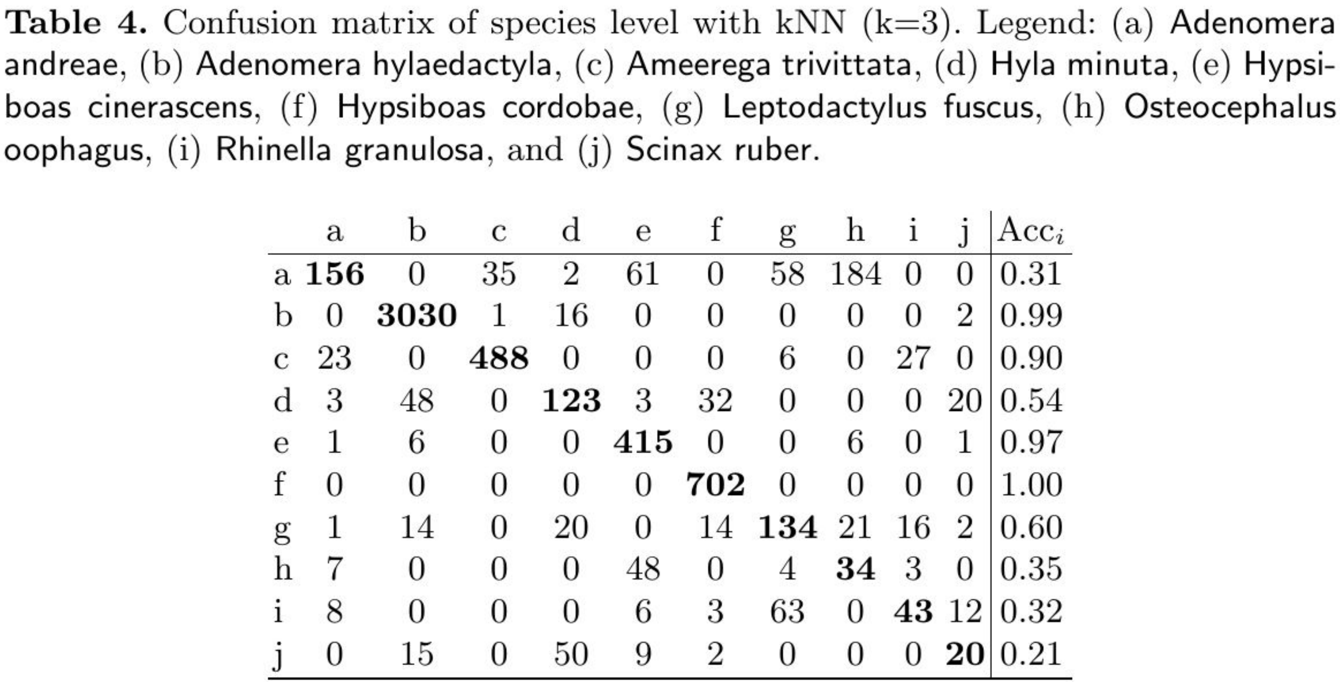


Table 4. Confusion matrix of species level with kNN (k=3). Legend: (a) Adenomera andreae, (b) Adenomera hylaedactyla, (c) Ameerega trivittata, (d) Hyla minuta, (e) Hypsiboas cinerascens, (f) Hypsiboas cordobae, (g) Leptodactylus fuscus, (h) Osteocephalus oophagus, (i) Rhinella granulosa, and (j) Scinax ruber.

	a	b	c	d	e	f	g	h	i	j	Acc_i
a	**156**	0	35	2	61	0	58	184	0	0	0.31
b	0	**3030**	1	16	0	0	0	0	0	2	0.99
c	23	0	**488**	0	0	0	6	0	27	0	0.90
d	3	48	0	**123**	3	32	0	0	0	20	0.54
e	1	6	0	0	**415**	0	0	6	0	1	0.97
f	0	0	0	0	0	**702**	0	0	0	0	1.00
g	1	14	0	20	0	14	**134**	21	16	2	0.60
h	7	0	0	0	48	0	4	**34**	3	0	0.35
i	8	0	0	0	6	3	63	0	**43**	12	0.32
j	0	15	0	50	9	2	0	0	0	**20**	0.21

Summary and conclusions

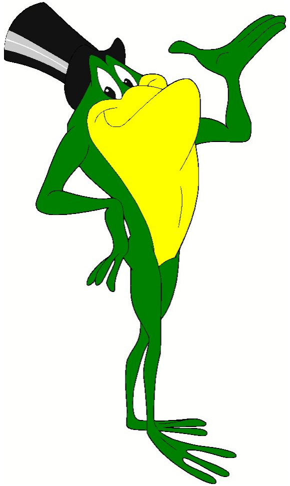
- **The hierarchical approach effectively reduces the complexity** of problems maintaining an acceptable accuracy.
- From a classification point of view the families Bufo, Hyla and Lepto were the most similar in the feature space, and also the species Adenomera andreae and Osteocephalus oophagus.
- The Scinax species was the most difficult to recognize.
- The ***k*****CV by individuals** (specimens) has an important impact in the model performance.
- Baseline comparison against a dummy random classifier:
  - Micro gain = +35% and Average gain = +50%

Future work: Implement soft decision rules in the tree to be able to correct the error propagation from the highest levels.

LIAAD
INESCTEC
LABORATÓRIO ASSOCIADO

U.PORTO

UNIVERSIDADE FEDERAL DO AMAZONAS
IN UNIVERSA SCIENTIA VERITAS



Thanks – Obrigado – Gracias

MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA