# Machine Learning in life sciences.

Ignacio Sánchez-Gendriz[1,2]

[1] Departamento de Engenharia de Computação e Automação (DCA/UFRN)

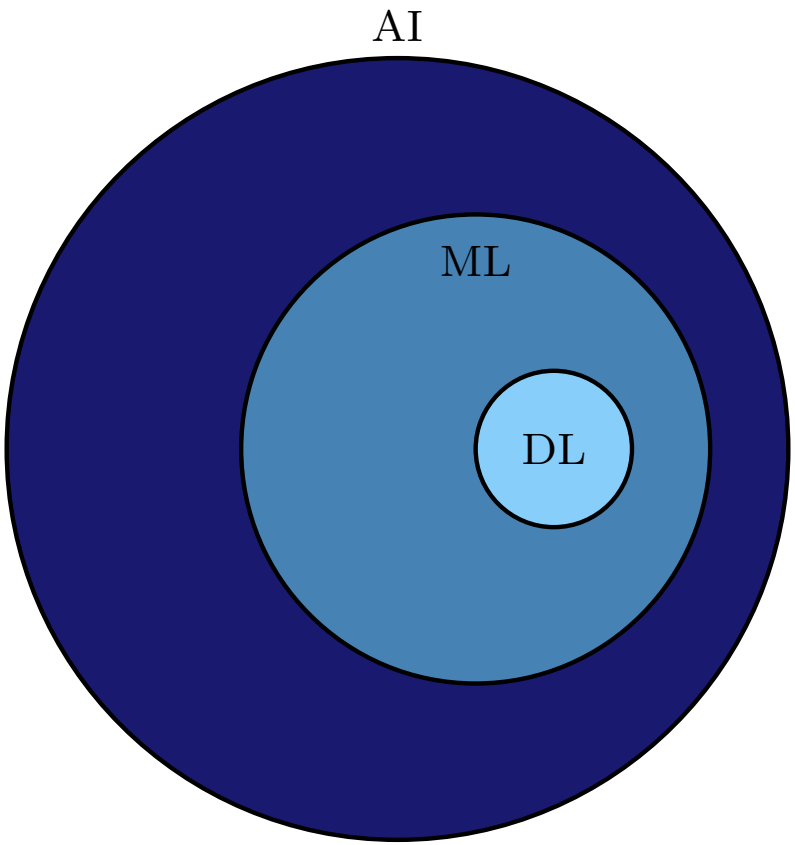
September 8, 2024

# Summary

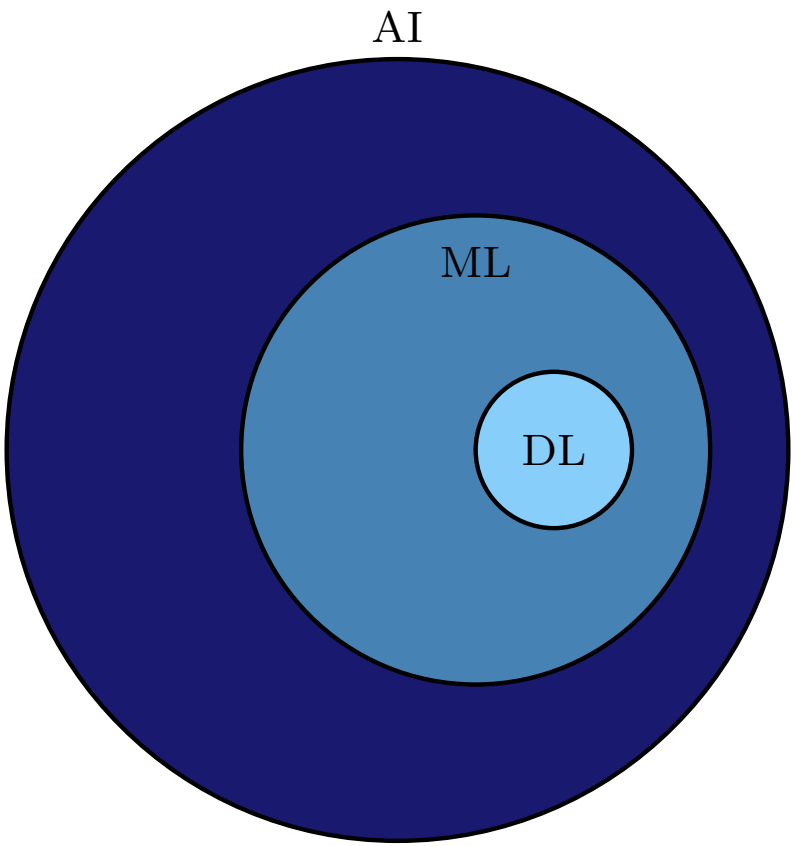


PSICO BIO TECNO LOGIA

# AI, ML & DL



# AI, ML & DL



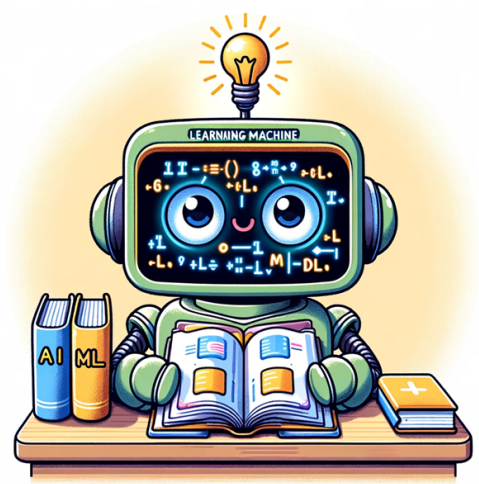

**Artificial Intelligence (AI):** Simulates intelligent behavior, akin to how animals adapt using rules and experiences.

**Machine Learning (ML):** Just as animals learn from experiences, ML trains computers with data to recognize patterns and make decisions.

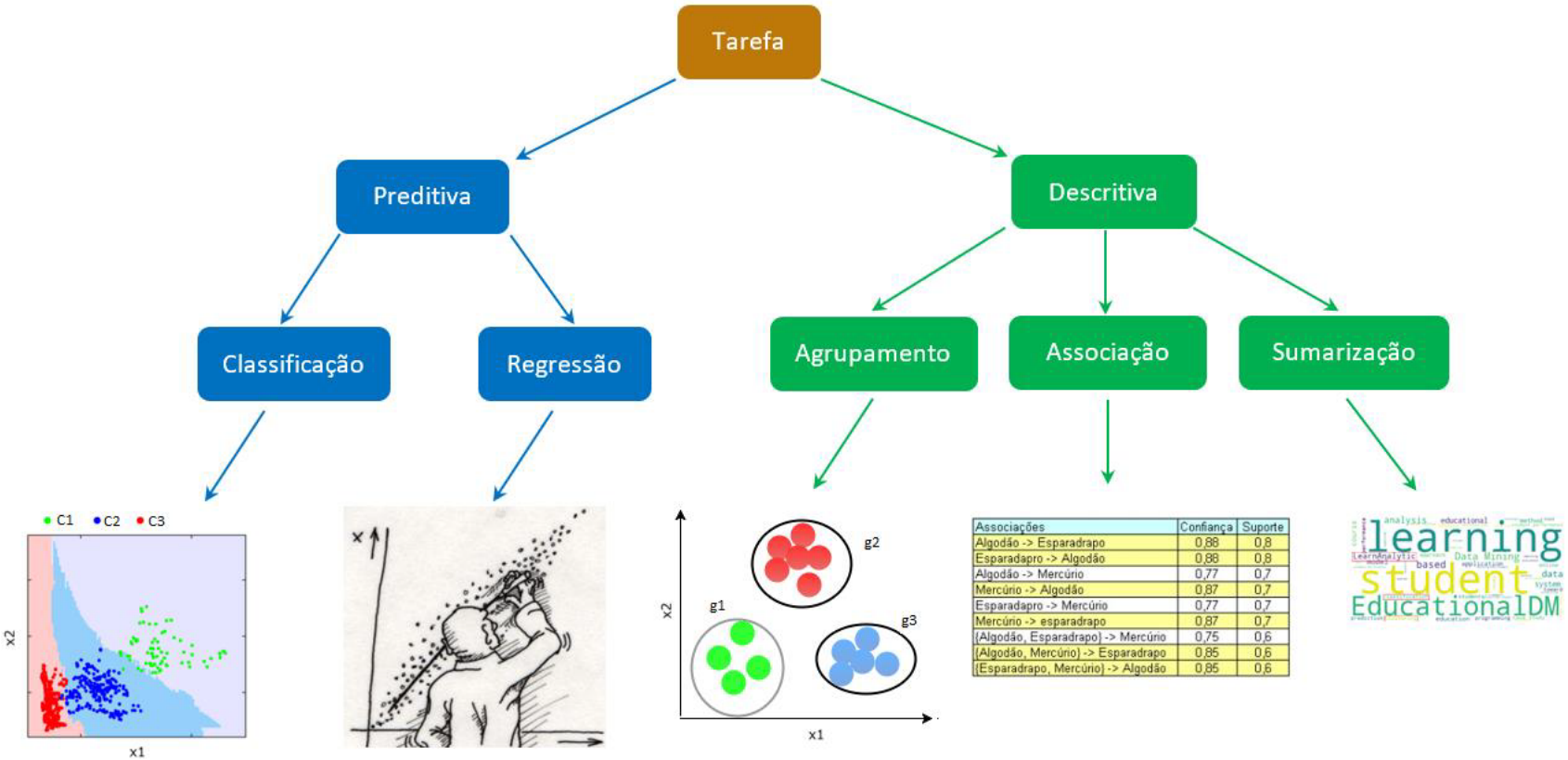
**Deep Learning (DL):** Like sifting through complex thoughts, DL allows computers to abstract and prioritize information, inspired by the depth of neural processes.

# ML Fundamentals

- **Machine Learning Models:** Trained using examples chosen for a specific task.
- **Training Data:** This consists of features that are derived from raw or preprocessed data, and are used to represent the examples during the model training phase.
- **Goal:** To discover patterns within the training examples that help generalize to new, unseen data, akin to discerning behavioral patterns that contextualize or predict future behaviors.

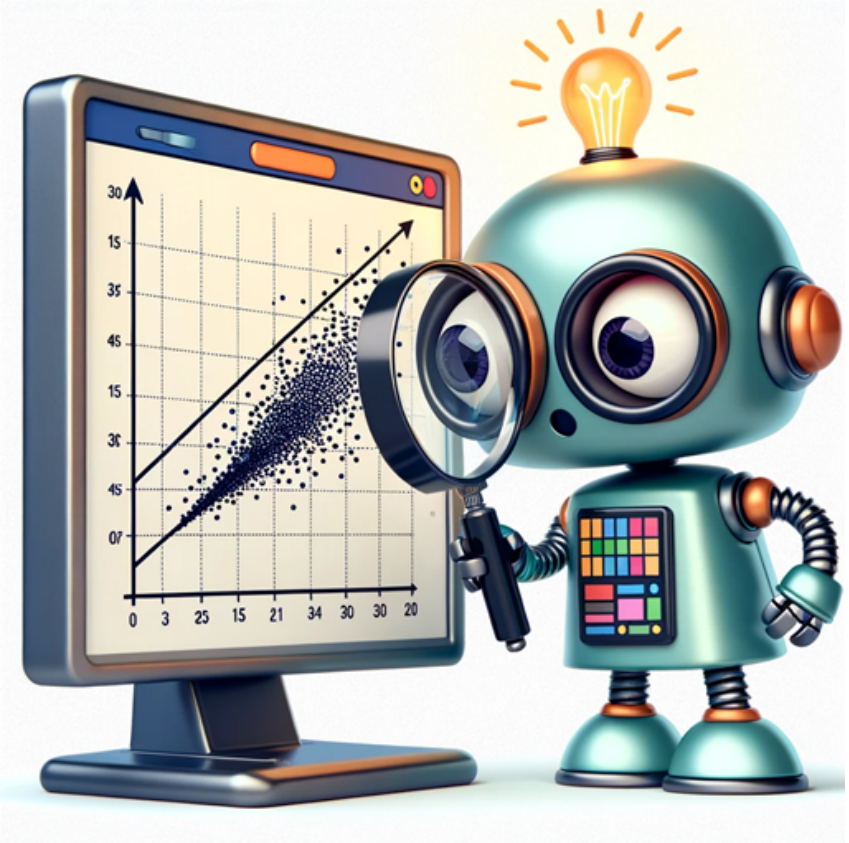


# ML Tasks


Tarefa
Preditiva
Descritiva
Classificação
Regressão
Agrupamento
Associação
Sumarização
C1 C2 C3
x1
x2
x
g1
g2
g3
x1
x2
Associações | Confiança | Suporte
Algodão -> Esparadrapo | 0,88 | 0,8
Esparadapro -> Algodão | 0,88 | 0,8
Algodão -> Mercúrio | 0,77 | 0,7
Mercúrio -> Algodão | 0,87 | 0,7
Esparadapro -> Mercúrio | 0,77 | 0,7
Mercúrio -> esparadrapo | 0,87 | 0,7
{Algodão, Esparadrapo} -> Mercúrio | 0,75 | 0,6
{Algodão, Mercúrio} -> Esparadrapo | 0,85 | 0,6
{Esparadrapo, Mercúrio} -> Algodão | 0,85 | 0,6
learning
student
EducationalDM

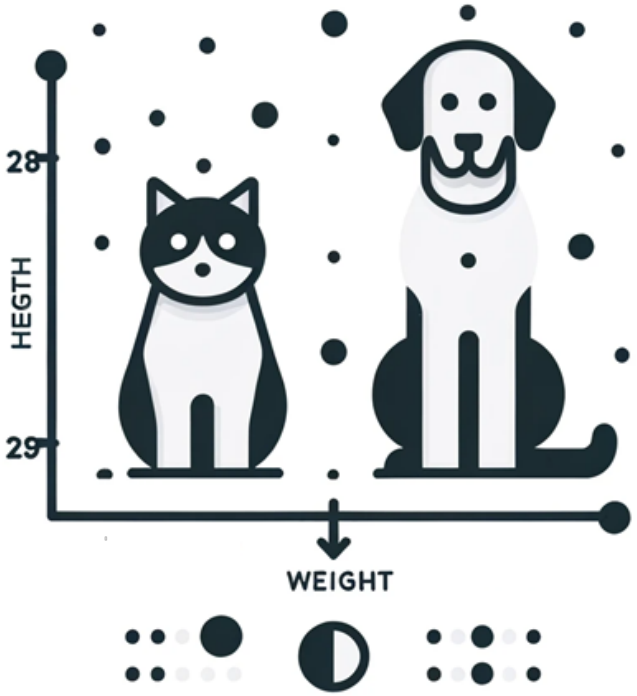
# The relevance of Visualization and EDA

*A **good** picture is worth a thousand words*, which is particularly true for data analysis.



# Representation of Objects

Objects are represented in the ML world by their features, which can be understood as variables measured from the particular object.



# Distance and Similarity

Distance in the context of ML is a way to quantify how different two points are from each other. It is like measuring how far apart two objects are, but instead of using a ruler, we use mathematical functions to consider their features or characteristics.

Euclidean distance is the "ordinary" straight-line distance between two points. For two points $\mathbf{p}$ and $\mathbf{q}$ in an 2D space is:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

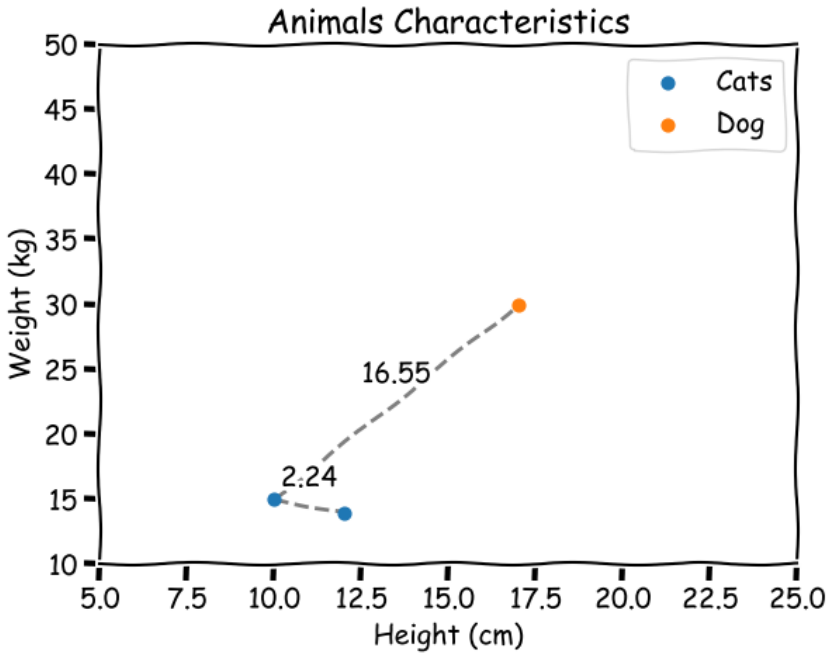



Figure: Caption

# Dimensionality Reduction

Data can have many characteristics, some of which may be irrelevant or contain redundant information. ML models may perform poorly with high-dimensional data (too many variables) due to increased processing time and memory resources (forced to deal with many variables). To address this, Dimensionality Reduction techniques can be applied, such as **Feature Selection** and **Feature Extraction**.

# Feature Selection

Feature selection is a process used to select relevant features for a specific analytical task. It simplifies models, improves performance, and provides clearer insights by removing unnecessary data. These methods range from automated machine learning techniques to expert-driven heuristics.

# Feature Selection

Now, lets imagine the following context: We conduct a memorization experiment with zebrafish in a laboratory setting, during which we collect Behavioral Data consisting of 12 Features. How should we analyze this data?

| Behavioral Data Features[1] | | | |
|---|---|---|---|
| Group | Animal ID | MemoryObjA | MemoryObjA |
| DiscriminatingObjA | DiscriminatingObjB | **MemoryVel** | **DiscriminatingVel** |
| MemoryFreez | DiscriminatingFreez | **MemoryDTP** | **DiscriminatingDTP** |

[1] Data Collected at https://www.luchiarilab.com/

# Feature Selection

Now, lets imagine the following context: We conduct a memorization experiment with zebrafish in a laboratory setting, during which we collect Behavioral Data consisting of 12 Features. How should we analyze this data?

| Behavioral Data Features[1] | | | |
|---|---|---|---|
| Group | Animal ID | MemoryObjA | MemoryObjA |
| DiscriminatingObjA | DiscriminatingObjB | **MemoryVel** | **DiscriminatingVel** |
| MemoryFreez | DiscriminatingFreez | **MemoryDTP** | **DiscriminatingDTP** |

For instance, we could select four relevant features based on 'specialist' criteria, which might come from a human expert or an ML model.

[1] Data Collected at https://www.luchiarilab.com/

# Feature Extraction

”Sometimes, the feature selection approach may not be the best option, or it may be part of a more complex task. Suppose we want to visualize the data from the four selected features mentioned earlier. How should we proceed in that case?

Table: Fish Behavioral Sample Data

| Fish ID | Group | MemoryVEL | DiscriminatingVEL | MemoryDTP | DiscriminatingDTP |
|---|---|---|---|---|---|
| #1 | 6 months | 0.777 | 0.000 | 0.777 | 0.000 |
| #3 | 6 months | 0.587 | 0.778 | 0.587 | 0.778 |
| #5 | 6 months | 0.728 | 0.396 | 0.728 | 0.396 |
| #7 | 6 months | 0.700 | 0.944 | 0.700 | 0.944 |
| #9 | 6 months | 1.000 | 0.472 | 1.000 | 0.472 |
| #11 | 18 months | 0.484 | 0.450 | 0.484 | 0.450 |
| #13 | 19 months | 0.000 | 0.557 | 0.000 | 0.558 |
| #15 | 21 months | 0.342 | 0.544 | 0.341 | 0.544 |
| #17 | 23 months | 0.291 | 0.711 | 0.290 | 0.711 |
| #19 | 25 months | 0.230 | 0.469 | 0.230 | 0.469 |

# Feature Extraction

”Sometimes, the feature selection approach may not be the best option, or it may be part of a more complex task. Suppose we want to visualize the data from the four selected features mentioned earlier. How should we proceed in that case?

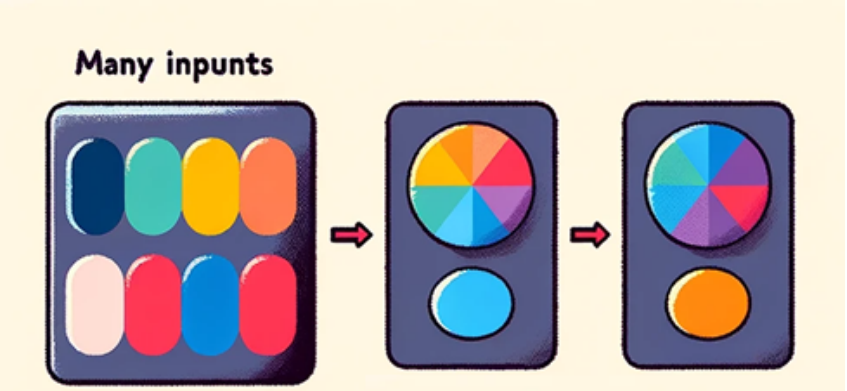
Table: Fish Behavioral Sample Data

| Fish ID | Group | MemoryVEL | DiscriminatingVEL | MemoryDTP | DiscriminatingDTP |
|---|---|---|---|---|---|
| #1 | 6 months | 0.777 | 0.000 | 0.777 | 0.000 |
| #3 | 6 months | 0.587 | 0.778 | 0.587 | 0.778 |
| #5 | 6 months | 0.728 | 0.396 | 0.728 | 0.396 |
| #7 | 6 months | 0.700 | 0.944 | 0.700 | 0.944 |
| #9 | 6 months | 1.000 | 0.472 | 1.000 | 0.472 |
| #11 | 18 months | 0.484 | 0.450 | 0.484 | 0.450 |
| #13 | 19 months | 0.000 | 0.557 | 0.000 | 0.558 |
| #15 | 21 months | 0.342 | 0.544 | 0.341 | 0.544 |
| #17 | 23 months | 0.291 | 0.711 | 0.290 | 0.711 |
| #19 | 25 months | 0.230 | 0.469 | 0.230 | 0.469 |

One possible option is to apply feature extraction, but what exactly does that entail?

# Feature Extraction - PCA

Feature extraction can be understood as the process of combining the original features to generate new ones, creating a new representation space. The aim is essentially to represent the data with fewer variables without losing significant information. One of the most widely applied techniques in this field is Principal Component Analysis (PCA).



# Feature Extraction - PCA

PCA:

- Transform the original variables into a new feature space through linear combinations.
- Ensure the variables in the new feature space are uncorrelated, meaning each represents unique information.
- Capture a significant portion of the original information within a reduced number of principal components, reflecting most of the variance from the original data.
- Facilitate the visualization of underlying patterns in the data by reducing its dimensionality.

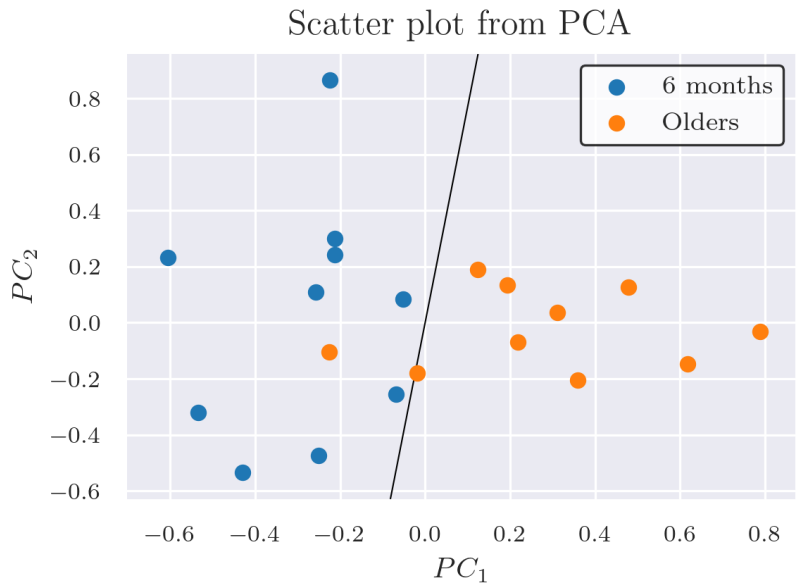


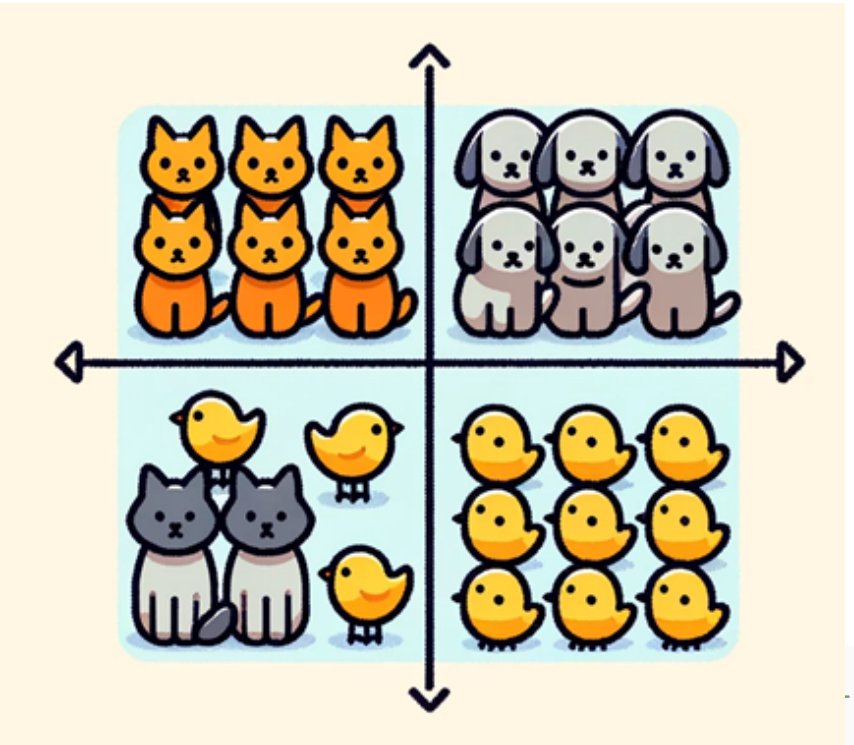

Figure: PCA of Fish Behavioral Data

# Clustering

Clustering techniques are considered unsupervised because they do not rely on pre-labeled data; there is no 'supervisor' to designate the category or class for each item. The primary goal of these methods is to group data such that objects within the same cluster have minimal distances between them, indicating similarity, while maximizing the distances between objects in different clusters to highlight their dissimilarity.

Popular clustering techniques include:

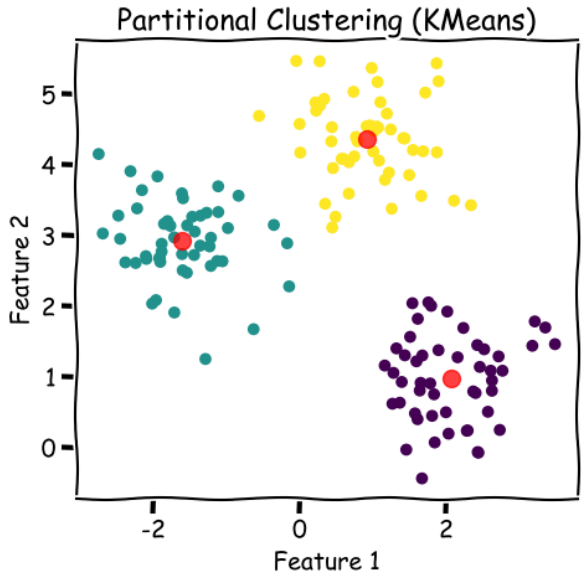
- Partitional: K-means algorithm segments data into K clusters by reducing in-cluster variance.
- Hierarchical: Dendrograms visualize data divisions based on nested group similarities.
- Density-Based: DBSCAN identifies and expands clusters from dense core samples, effectively differentiating high and low-density regions.



# K-Means Clustering

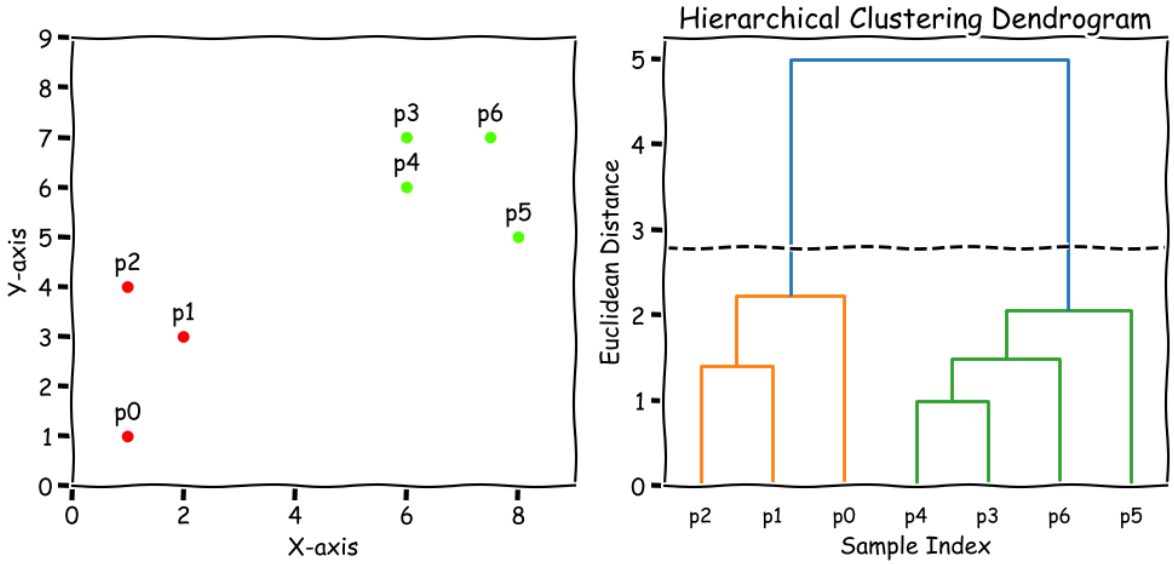
Due to its simplicity, it is one of the most utilized methods. Model Description:

1. Select $k$ centroids $c_i (i = 1, 2, , k)$.
2. Calculate the distance to each $c_i$.
3. Assign each instance to the group $G_i$ with the nearest centroid.
4. Recalculate the values of $c_i$ as the mean of their respective $G_i$.
5. Repeat steps 2 to 3 until there are no changes in the values of $c_i$.



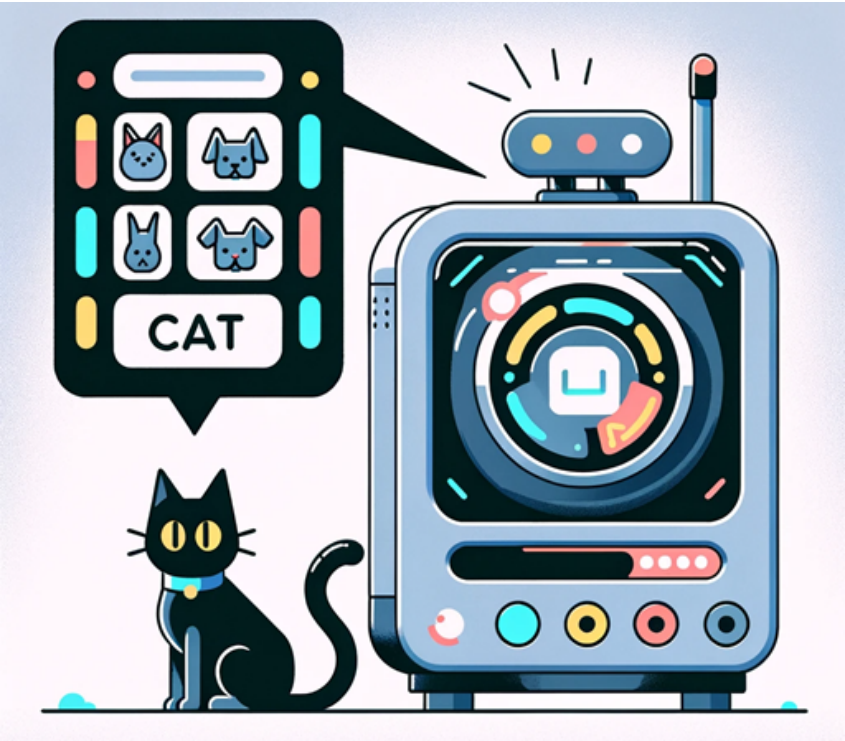
# Hierarchical Clustering

1. Starts with each instance in its own separate group.
2. Calculates all pairwise distances between the centroids of the groups.
3. Joins the two groups with the closest centroids.
4. Recalculates the centroid of the new merged group.
5. Repeats steps 2 to 4 until the desired number of clusters is achieved or a single cluster remains.



# Supervised ML - Classification

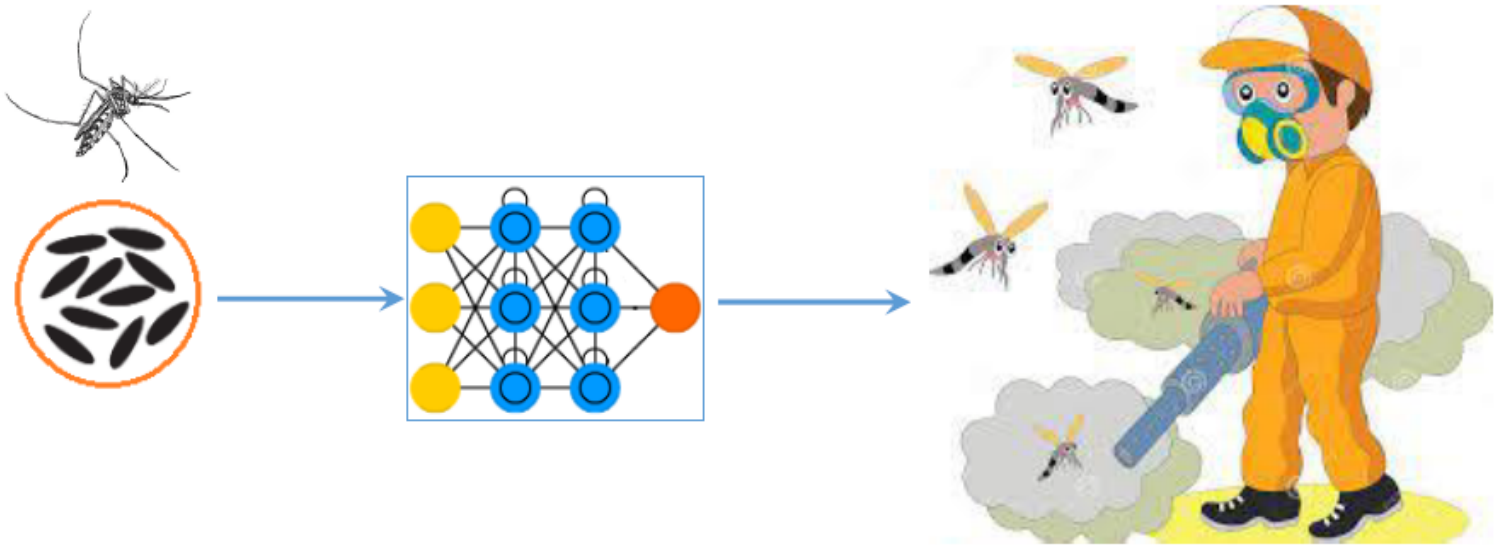
Classification stands as a prevalent task in machine learning where algorithms are trained using labeled data. This process justifies the 'supervised' learning tag, as it requires prior labeling of examples. The algorithm then discerns class patterns to predict the classification of new, unseen data.



”Let’s get our hands dirty with data and our minds clear for results”

# Ovitrap Data Analysis

This study aimed to train machine learning (ML) models to predict dengue case incidences,also guiding timely public health interventions for controlling the disease vector in Natal's city.



# Ovitrap Data Analysis

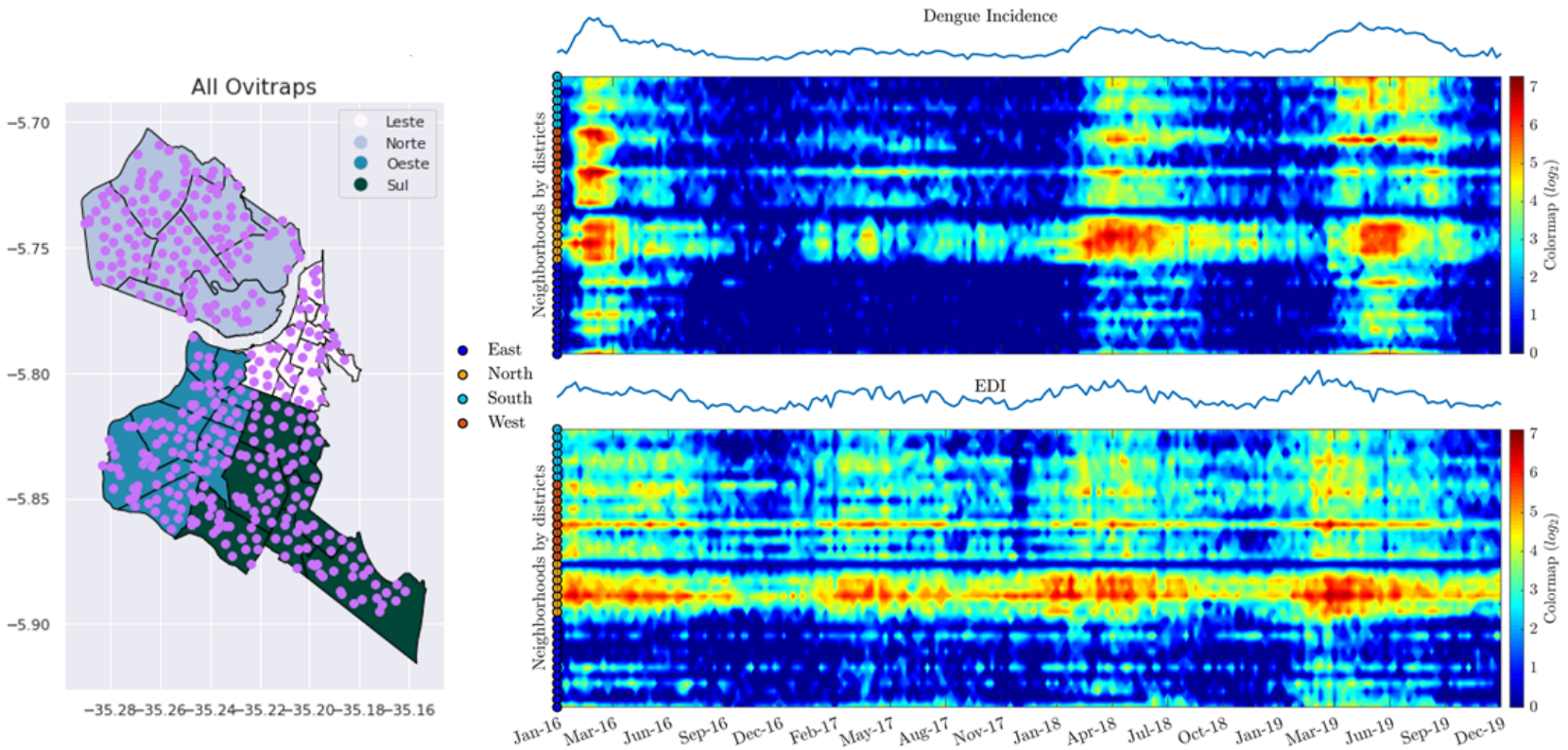



Figure: Visualization of Ovitrap Data: The left panel displays the distribution of ovitraps. The right panel depicts heatmaps representing the Egg Density Index (EDI) alongside dengue incidence rates by neighborhoods.

# EDI vs. Precipitation

Visual inspection corroborates that periods of elevated precipitation coincide with an increase in EDI.

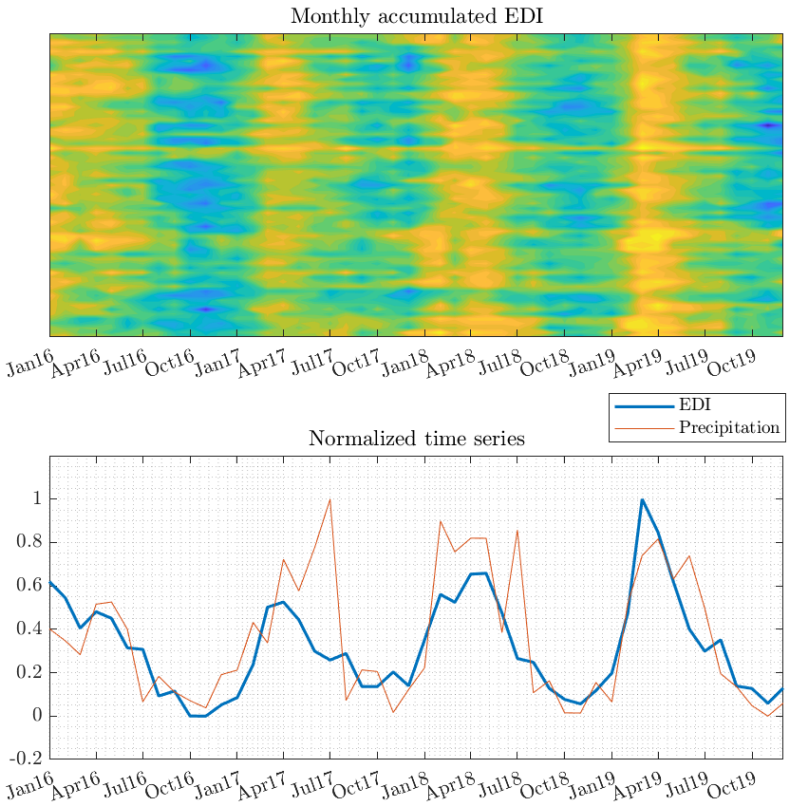



Figure: EDI and precipitation demonstrate a statistically significant and robust correlation ($r = 0.72$, p-value of $9.76 \times 10^{-9}$).

# Performance of models based on the mean EDI

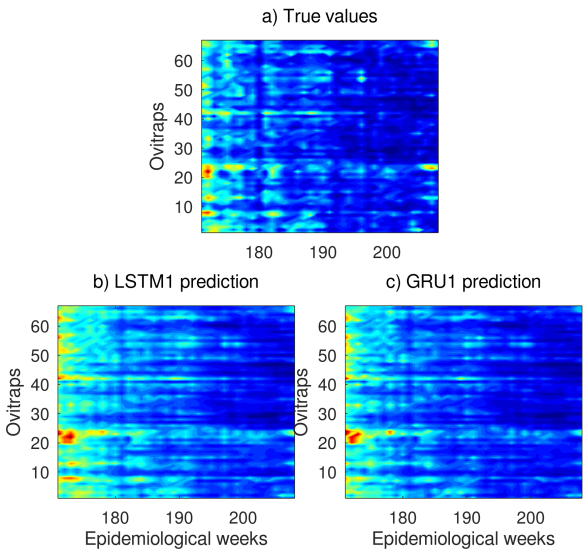



Figure: The models adequately reflected overall trends within individual dynamics.

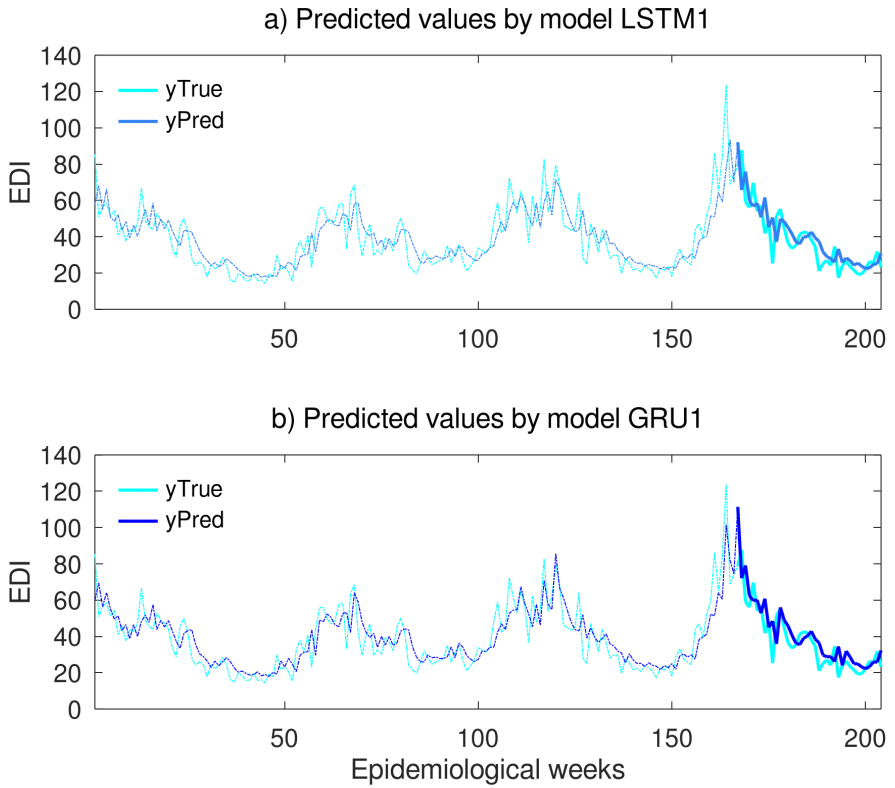



Figure: Two ML modles that reach 9.7 and 10.5 RMSE for predicting EDI

# Ovitrap Data - Final Remarks

- Our study underscores the applicability of forecasting arbovirus vector populations through recurrent ML models.
- Spatial smoothing and aggregation techniques, paired with the inclusion of temporal dependencies, demonstrated noteworthy potential in aiding public health interventions, specifically for Aedes aegypti control.
- Accurate and low-error forecasting of vector populations facilitates the enactment of proactive, precisely-timed interventions during elevated risk periods, thereby optimizing resources and interventions based on spatial trends.
- Future research intends to include additional climatic variables as predictors, aiming to enhance models' performance for fine-grained spatial regions.

# Exploring Fish Choruses from Daily Spectrograms

Far from being silent realms, oceans are teeming with species that utilize and navigate through acoustic communication, especially in an environment where light is scarce but sound travels more effectively than in terrestrial settings. Within this context, fish choruses stand out as one of nature's most interesting phenomena.

Visualizing and interpreting large datasets of underwater sound is a complex task that demands extensive computational resources and human expertise.

# Exploring Fish Choruses from Daily Spectrograms

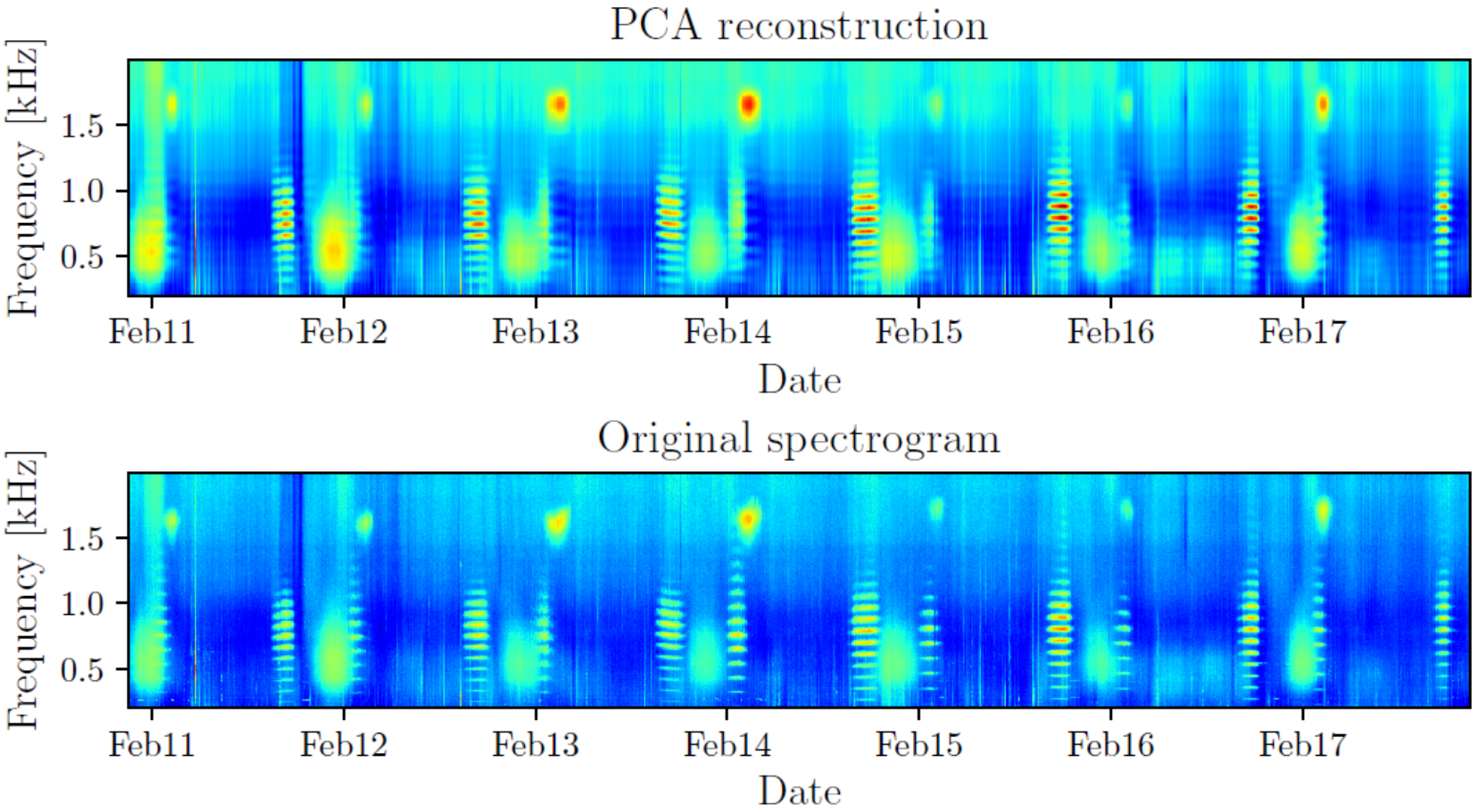
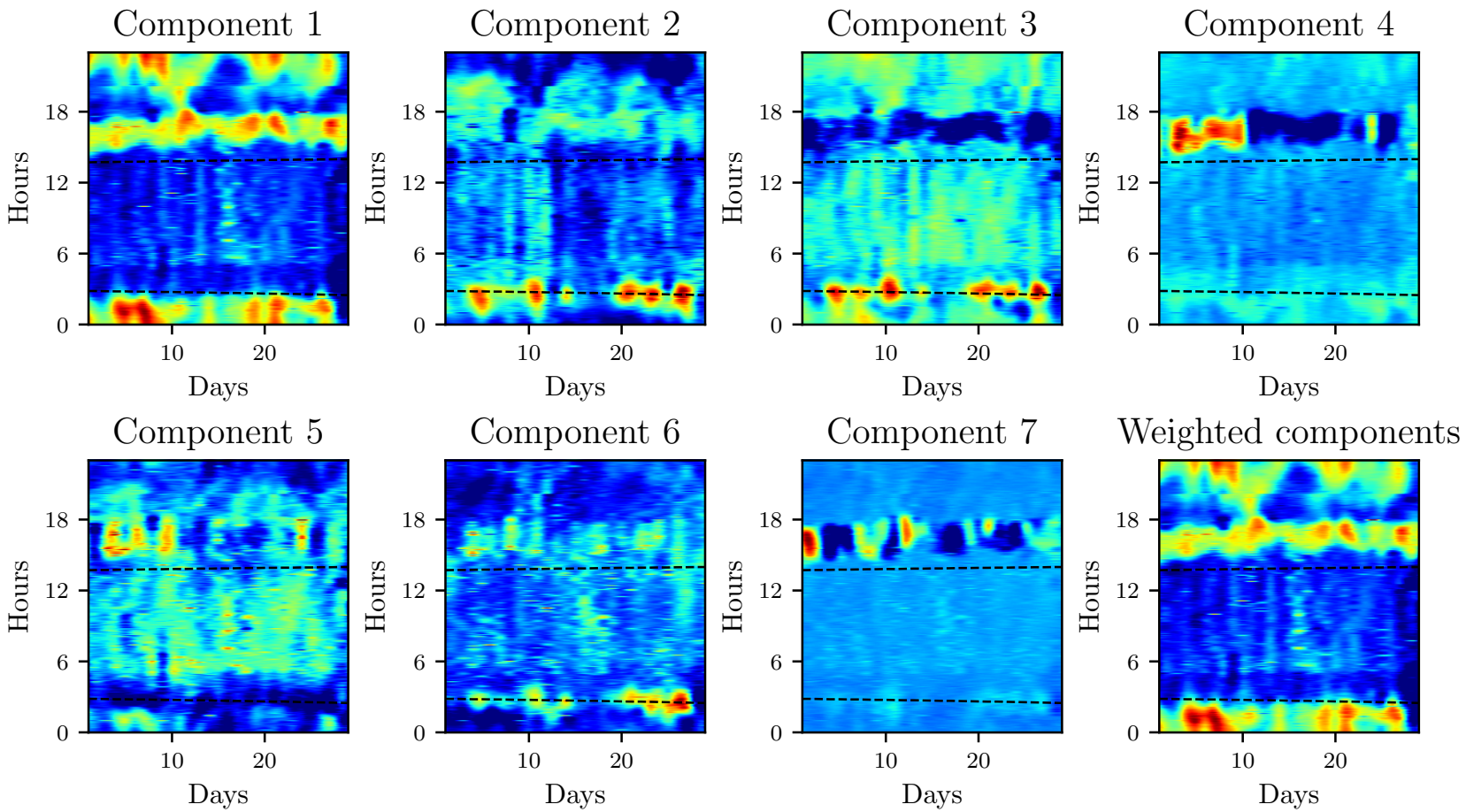



Figure: Weekly spectrogram comparison illustrating original data (bottom) and the reconstruction from PCA (top), emphasizing the effectiveness in capturing key acoustic patterns after PCA dimensionality reduction.

# Exploring Fish Choruses from Daily Spectrograms

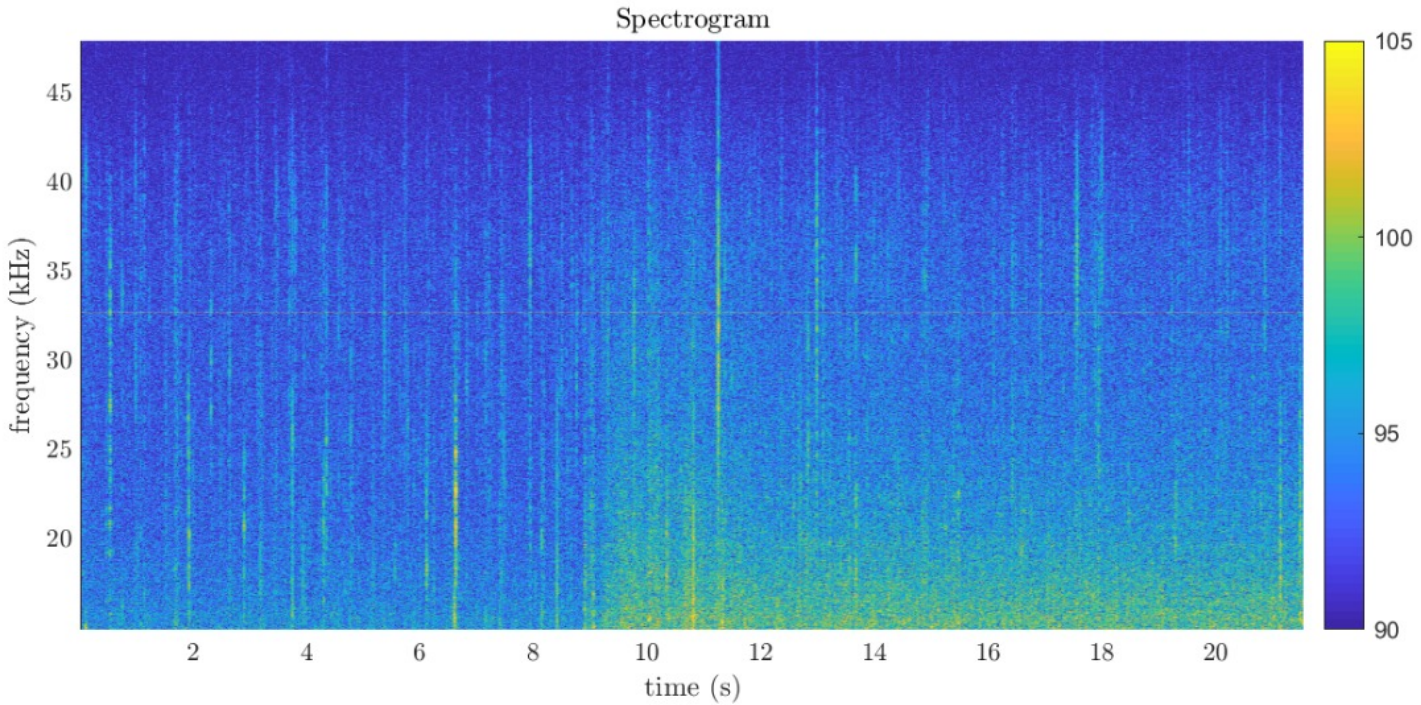
Depicting the Principal Components as heatmaps reveals variations in the studied soundscape, demonstrating correlations with sunrise and sunset times.



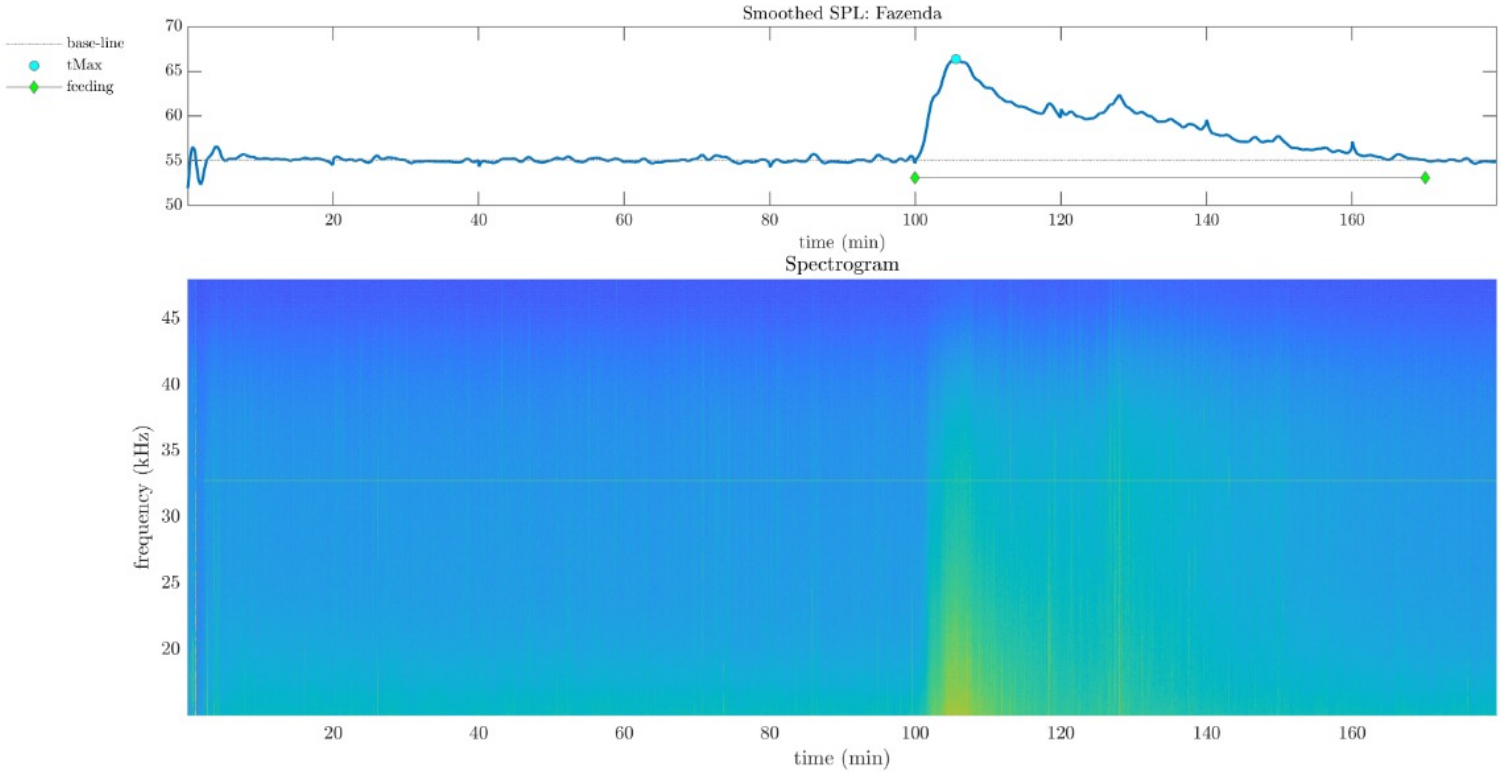
# Shrimp Feeding sound Analysis

Monitoring and controlling feeding activity in shrimp farms is critical for ensuring efficiency and promoting the well-being of animals in aquaculture. Given that shrimp feeding activity generates distinct impulsive sounds, this behavior can be effectively monitored through meticulous sound analysis.

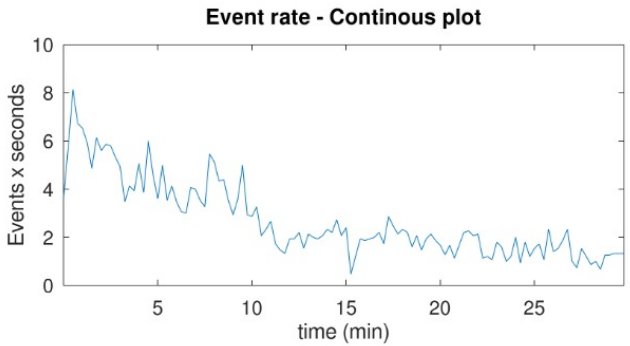
# Shrimp Feeding sound Analysis

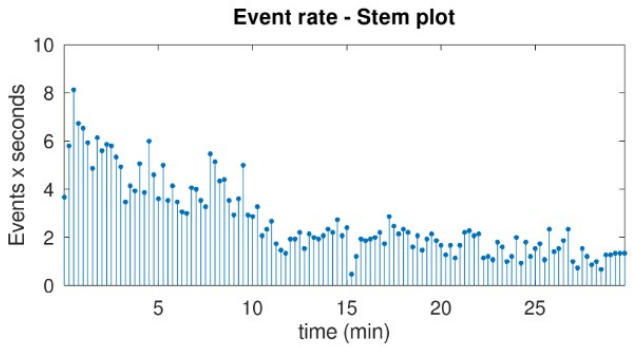


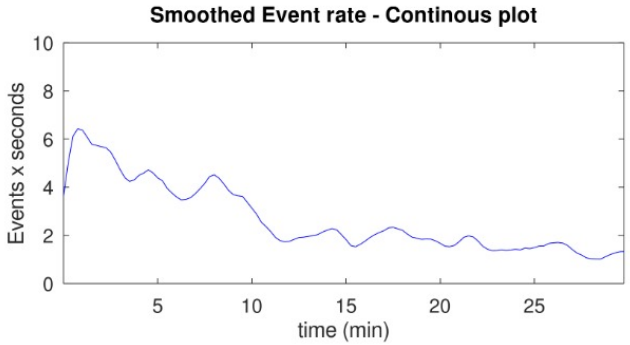
# Shrimp Feeding sound Analysis

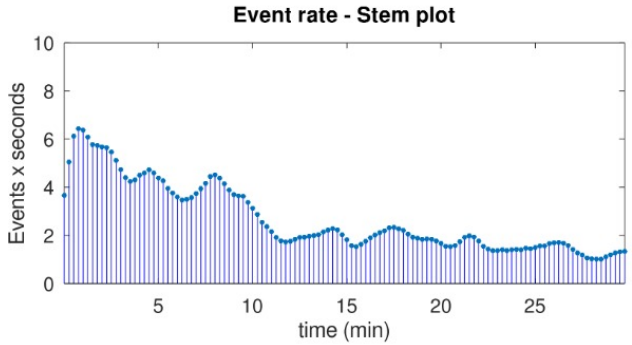


# Shrimp Feeding sound Analysis



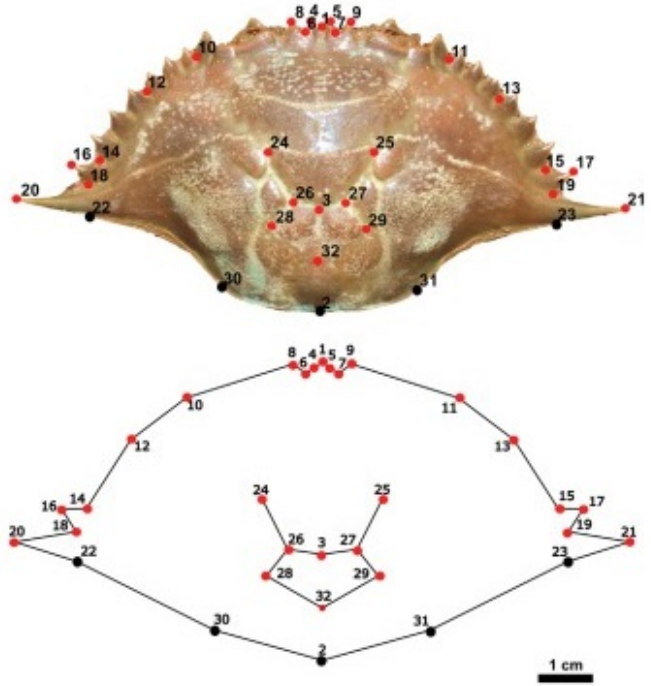



# Automatic Identification of Marine Species

Utilizing a dataset of fish images obtained in the field via the App Shiny4SelfReport,[2] and a controlled dataset, our research is focused on developing methods for the automatic detection of marine species by applying advanced Computer Vision and Machine Learning models.




Source: https://doi.org/10.1016/j.jcz.2021.09.009

[2] Supported by the TRIATLAS project and UFRN.
https://doi.org/10.1016/j.softx.2021.100843

# Cyclic Voltammetry Analysis for Dual Test Syphilis/HIV Detection

This project applies biosensors, signal processing, machine learning, and diagnostics to improve syphilis detection with CV sensors, offering a cost-effective diagnostic tool for STIs in primary health care.

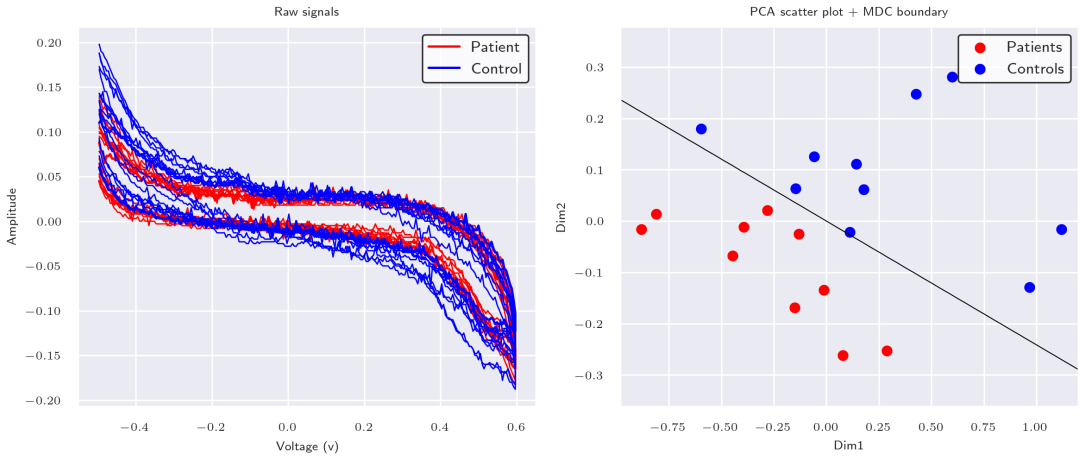



Table: Bootstrap cross-validation over 300 repetitions.

| Metric | Mean | Standard Deviation |
|---|---|---|
| Accuracy | 0.95 | 0.10 |
| Sensitivity | 0.99 | 0.08 |
| Specificity | 0.90 | 0.19 |
| F1 Score | 0.95 | 0.10 |

# DNA sequence analysis for virus classifications

The capacity to process large DNA sequences on standard computer hardware is critically important for quick responses to global health crises, like the COVID-19 pandemic.

Next, we presents the results of a new technique that translates DNA sequences into a two-dimensional numerical format. The findings illustrates the approach's effectiveness for classifying viruses and enabling visualizations with modest computational demands.

# DNA sequence analysis for virus classifications

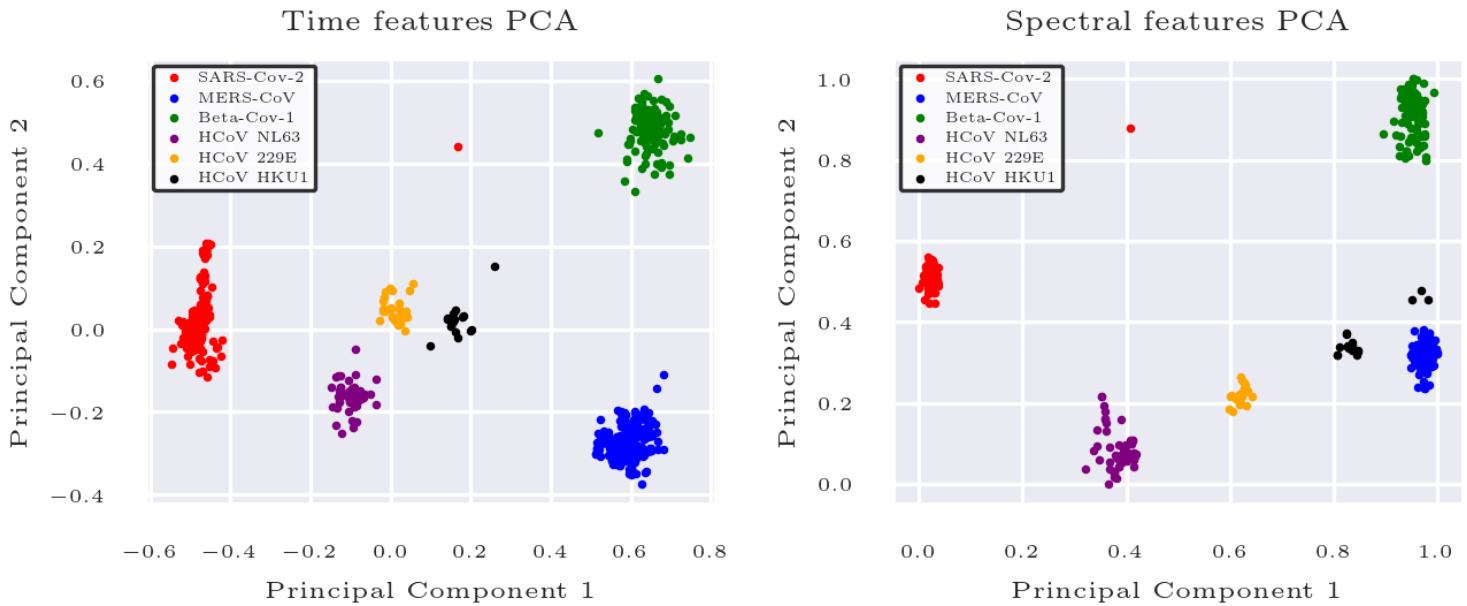



Figure: PCA Projection for Dimensionality Reduction. Panel A illustrates the application of PCA to the temporal domain features, demonstrating how data points are distributed in the principal component space. Panel B depicts the application of PCA to spectral domain features, showcasing the dispersion and grouping of data points in the reduced feature space.

# Conclusions

Throughout this talk, we've explored fundamental Machine Learning concepts and showcased their significant applicability within various life science domains. We've seen how interdisciplinary collaboration has become an irreversible and essential pathway, increasingly vital for sustaining high-quality research in academic institutions.

In my view, Machine Learning should not be perceived as a replacement for human researchers. Instead, it stands as a powerful tool that, when employed judiciously, can augment and amplify research capabilities within the life sciences and academia. It facilitates a symbiotic relationship where computational power and human insight combine to unravel the complex tapestries of biological phenomena.

Let us embrace Machine Learning as the robust and innovative ally it is, ensuring that we leverage its potential to enhance our understanding and contributions to the life sciences.

PSICO BIO TECNO LOGIA

# Acknowledgments

Thanks to all co-authors, collaborators, and supporters for their contributions to the research presented here. I also wish to acknowledge the organizers and supporters of the X Conference and the XV Symposium on Psychobiology for their role in fostering a vibrant scientific community.

Special thanks to Prof. Ricardo Valentim and team at LAIS/HUOL/UFRN for their collaboration and support.

My gratitude extends to the Project TRIATLAS team for their support.

Appreciation is due to Prof. Adrião Duarte Dória Neto, Prof. Luiz Affonso Guedes, and Prof. Marcelo Fernandes from DCA/UFRN for their invaluable insights and collaborations.

I am grateful to Fulvio Aurelio De Morais Freire for his steadfast support and collaboration.

Acknowledging the hard work and dedication of Matheus Diniz and Efrain Pulgar Pantaleon from DCA/UFRN.

Thanks to Prof. Silvio Peixoto and Dr. Santiago Hamilton from UFRPE for their collaborative spirit.

Thanks to Prof. Ana Carolina Luchiari for generously sharing data.

PSICO BIO TECNO LOGIA

# Questions?