Feature Selection of Long Non-Coding RNAs in Plants

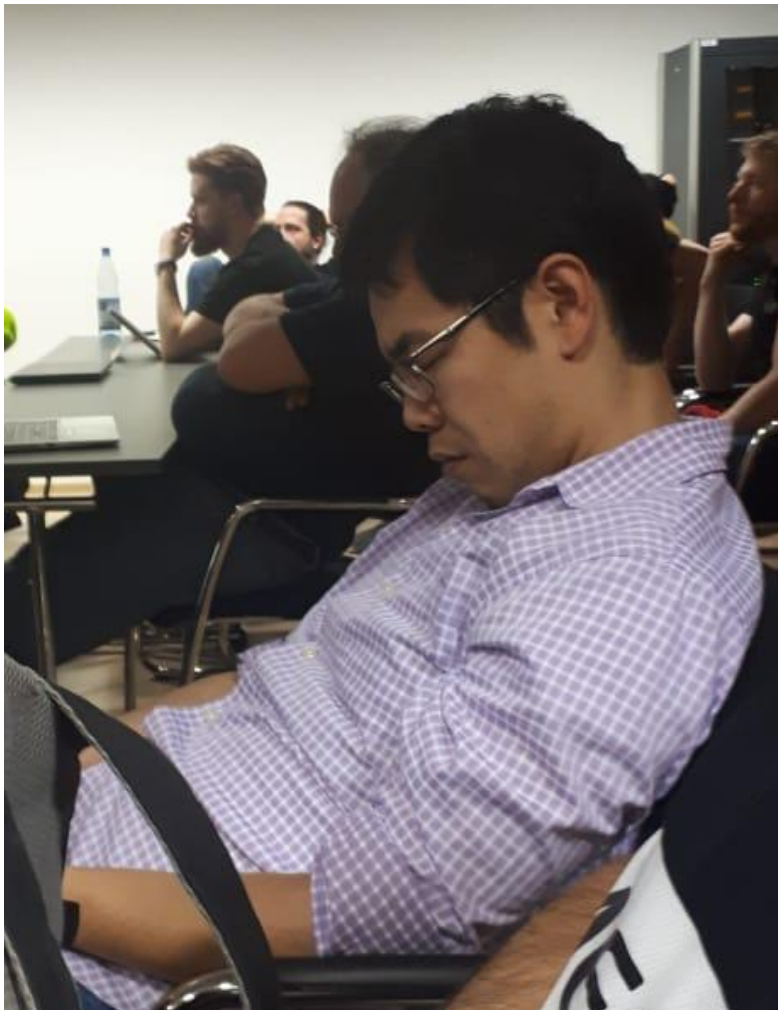
34th TBI Winterseminar in Bled - 14.02.19

Alexandre R. Paschoal

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Bioinformática UTFPR

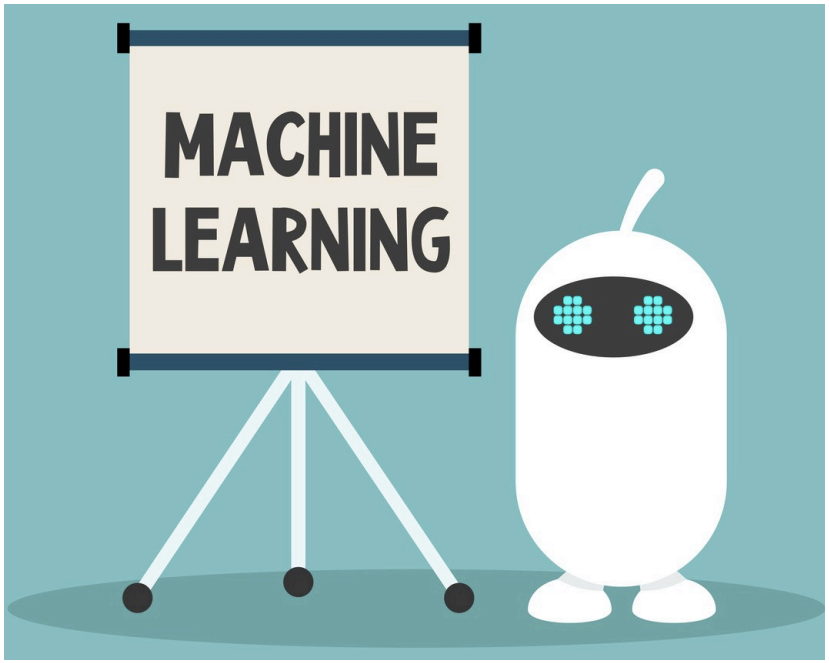
This talk is not about: Network ("learning") Statistics



This talk is not about: Network

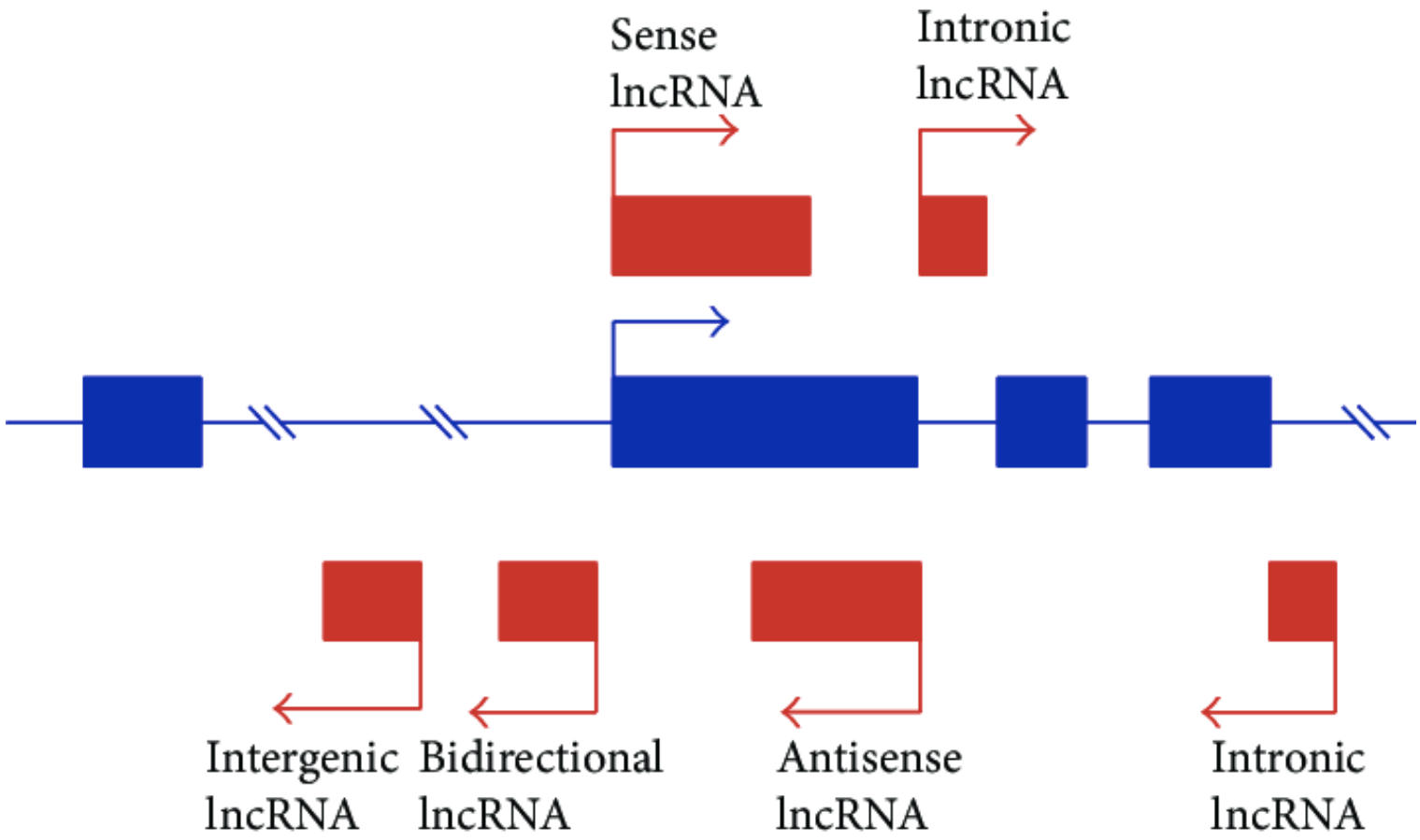
Motivation

- To understand which features contribute for long non-coding RNAs identification in plants.



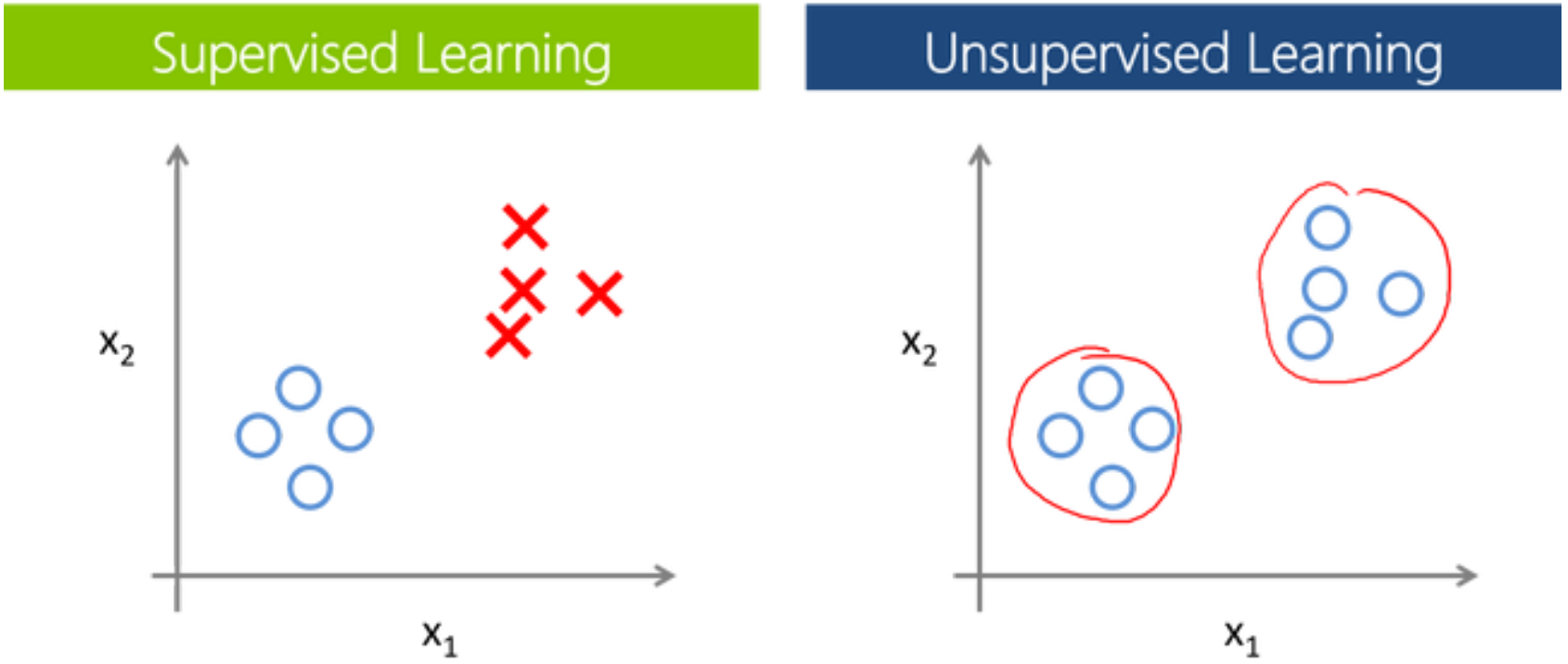
Definition: Long non-coding RNA

- Non-coding RNA > 200 nucleotide length



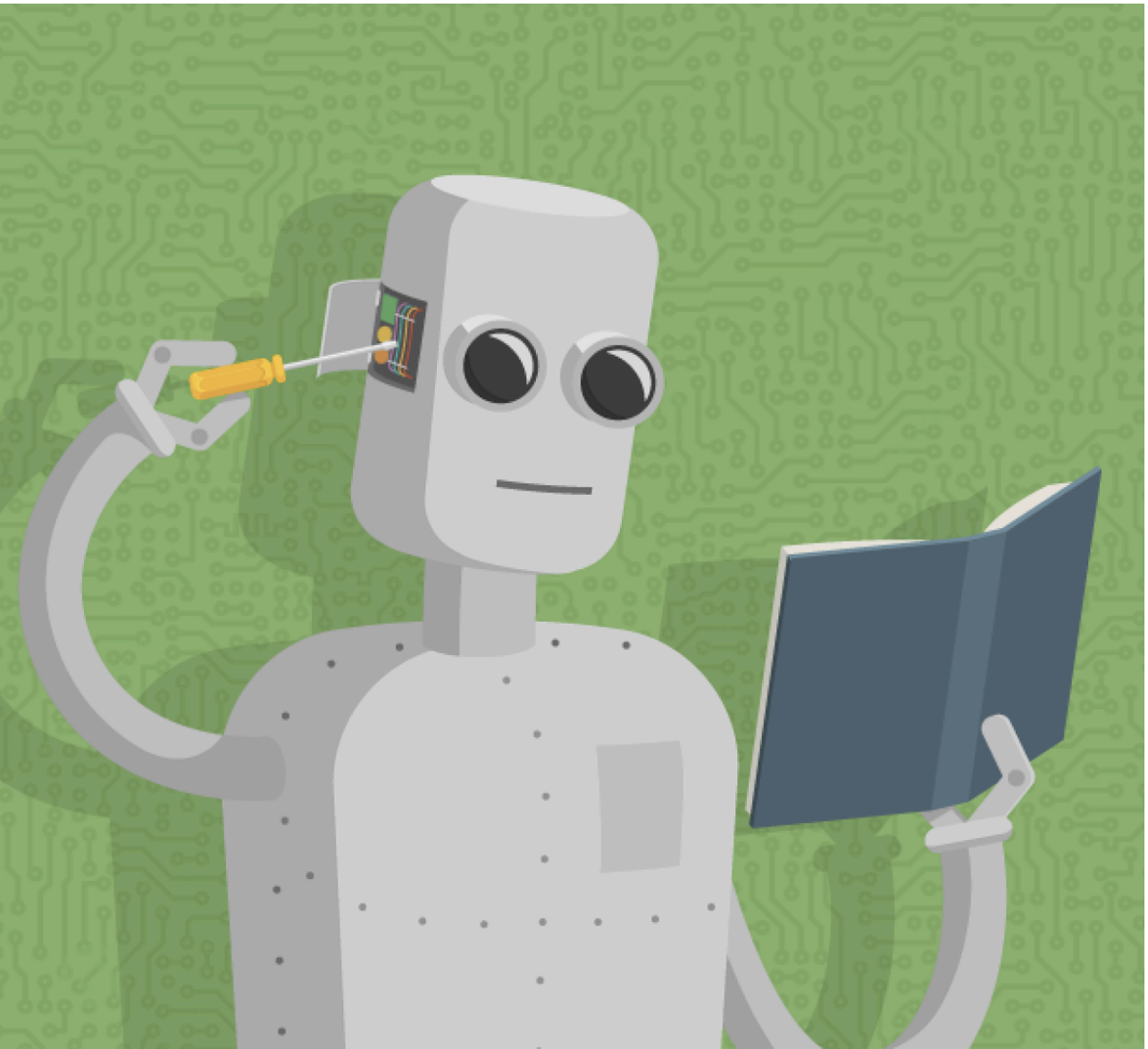
DOI: 10.1155/2016/5365209. 2016

ML: Supervised X Unsupervised Learning

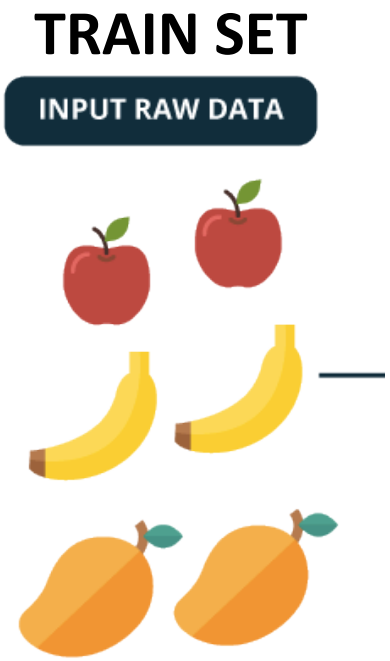


Source: https://lakshaysuri.wordpress.com/2017/03/19/machine-learning-supervised-vs-unsupervised-learning/

Machine Learning (ML)

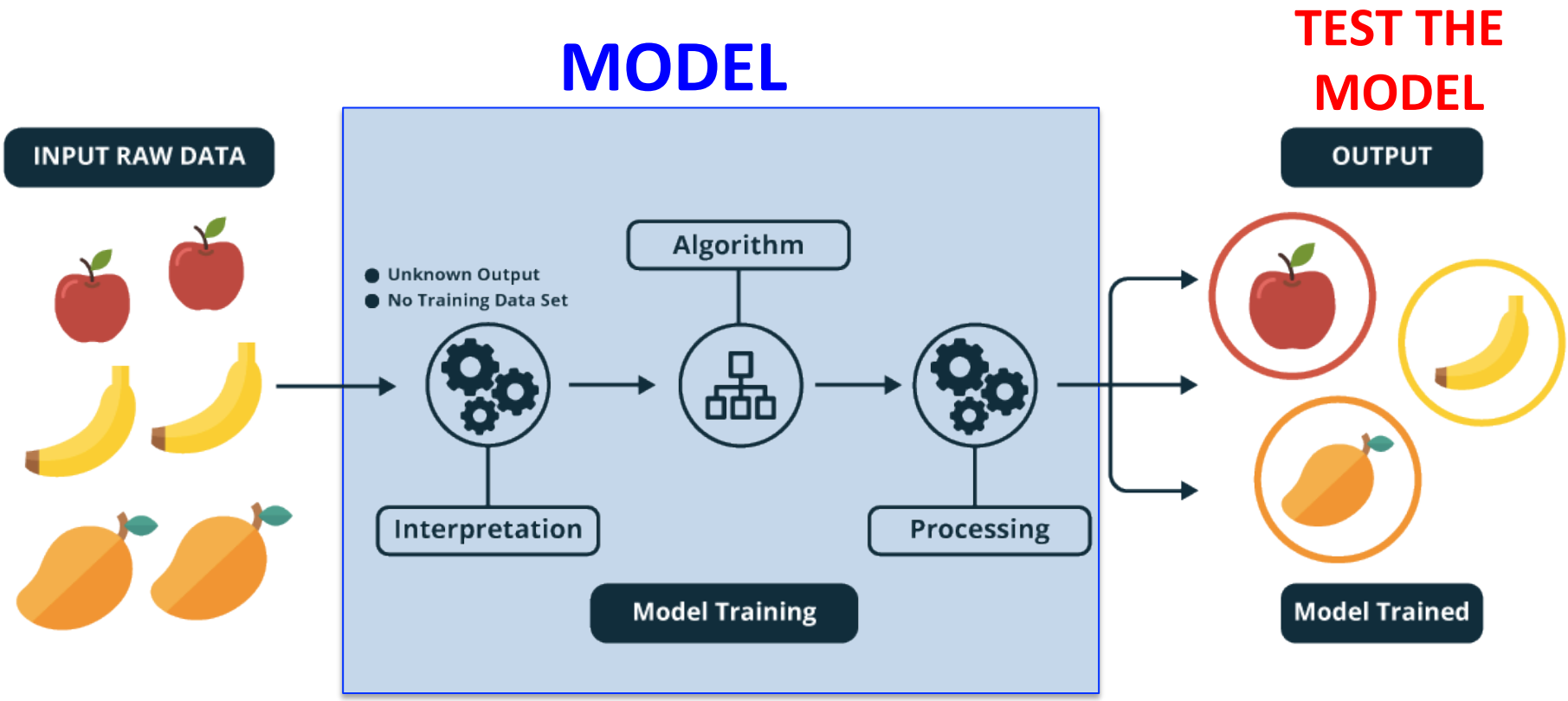


Machine Learning (ML)



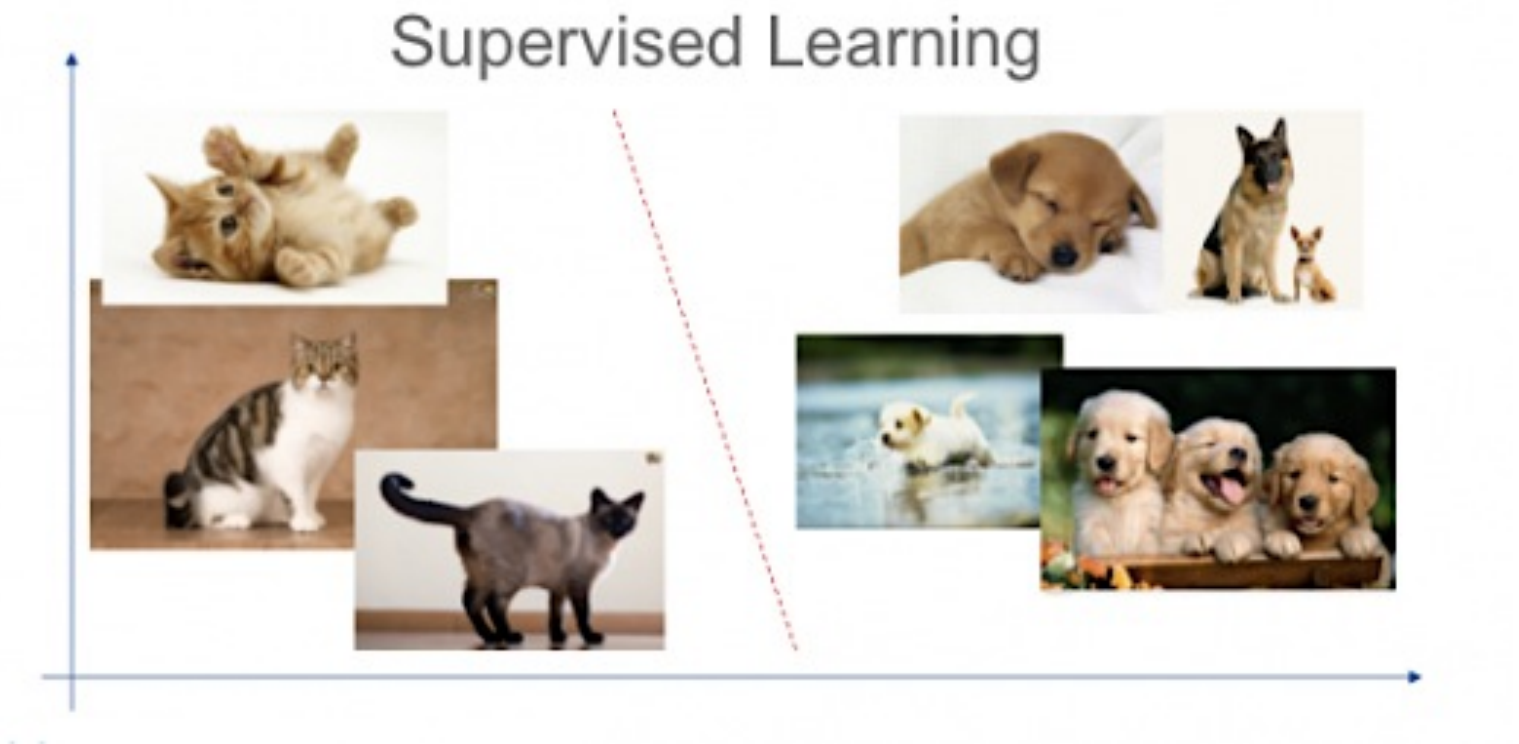
Source: https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Machine Learning (ML)



Source: https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Machine Learning (ML) - Example



Source: https://www.cisco.com/c/m/en_us/network-intelligence/service-provider/digital-transformation/get-to-know-machine-learning.html

To do the model

- **<u>Features to build the model</u>**:
  - Size of the fruit
  - Color the fruit
  - etc

Motivation

1. To use machine learning approaches for lncRNA identification

2. To investigate methods of feature selection

Previous work (Plants)

Briefings in Bioinformatics, 2018, 1–8

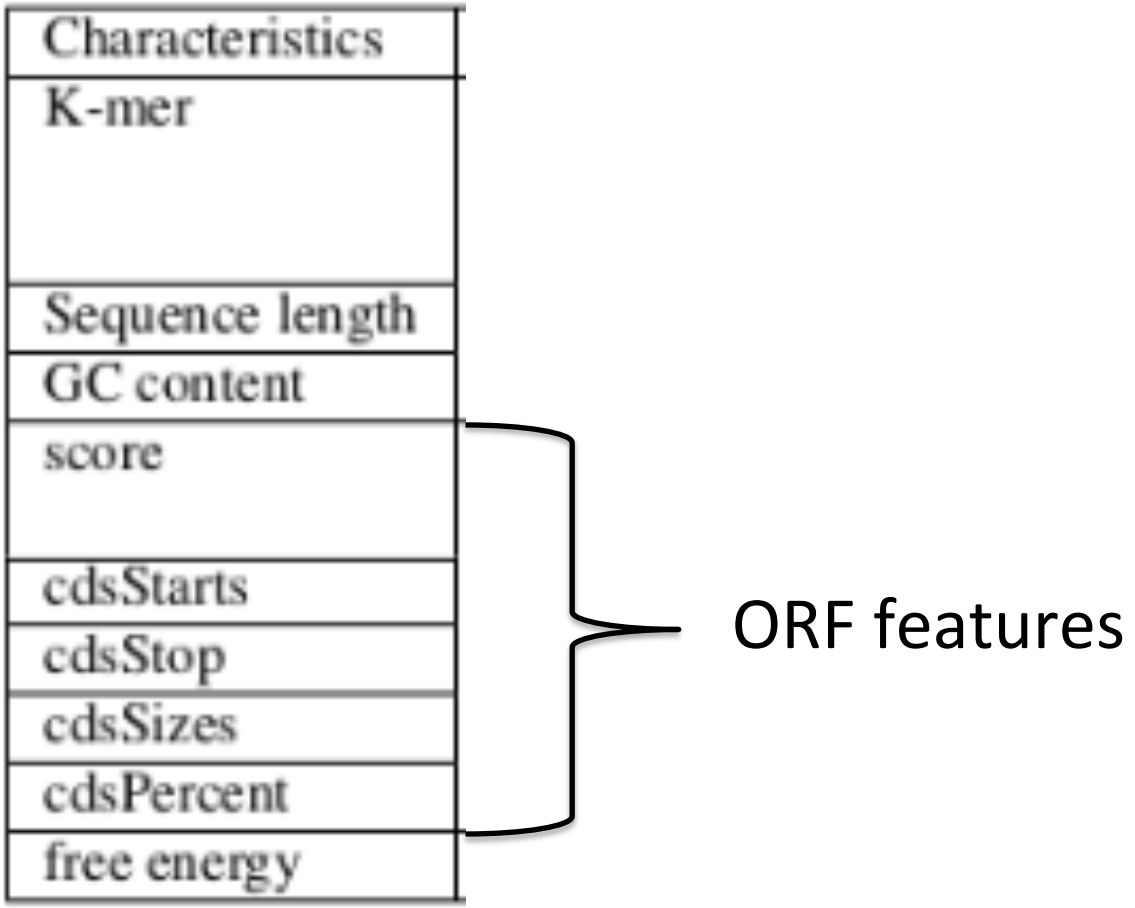
doi: 10.1093/bib/bby034
Review Paper

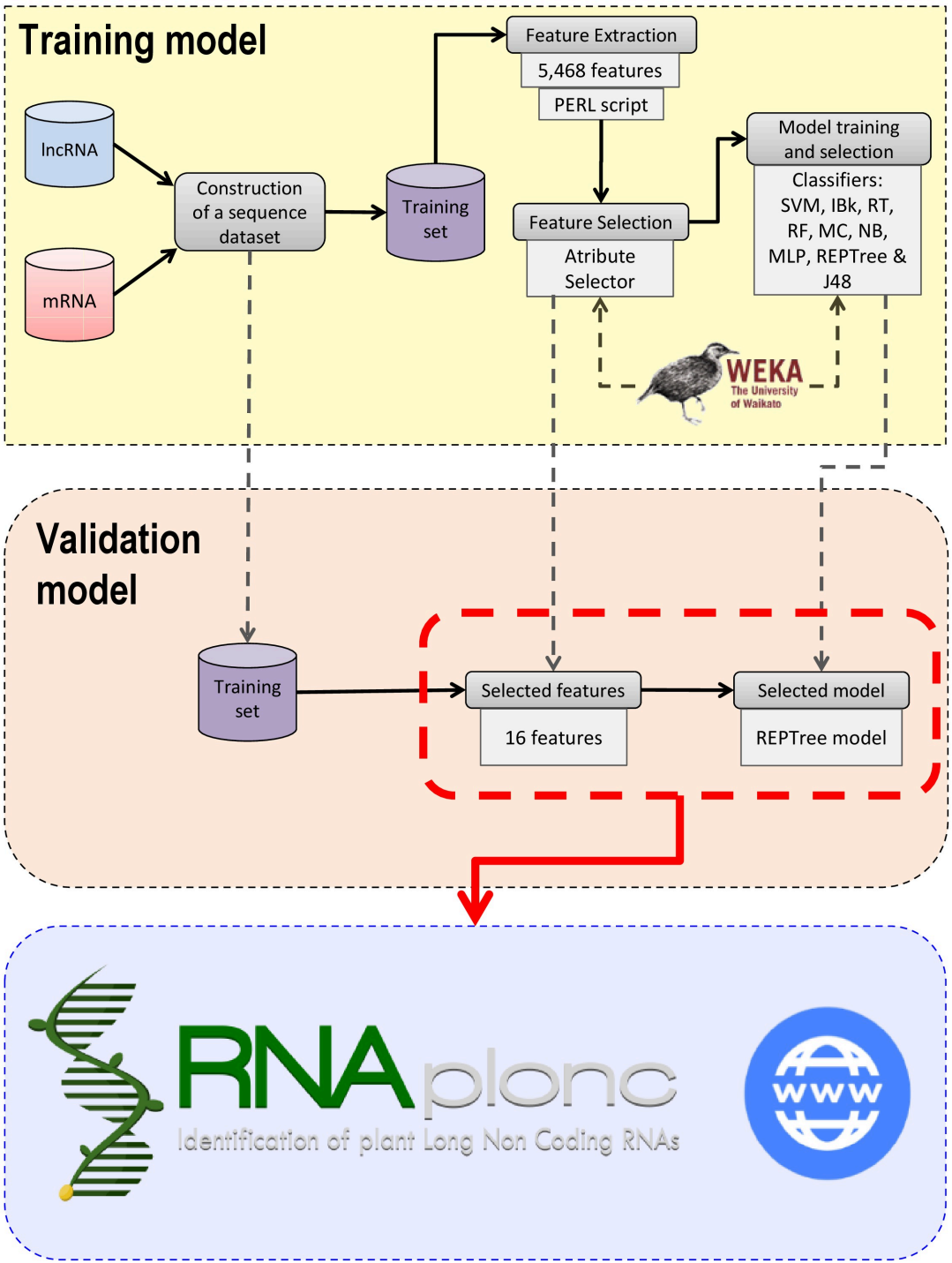
Pattern recognition analysis on long noncoding RNAs: a tool for prediction in plants

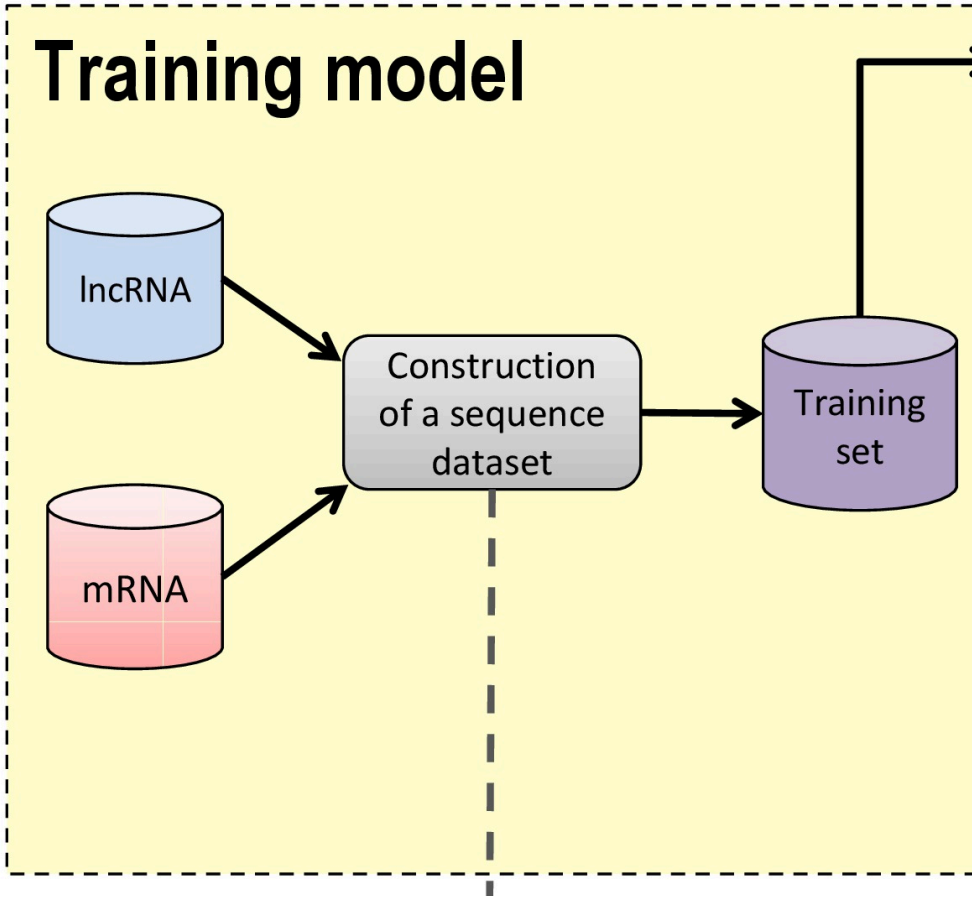
Tatianne da Costa Negri, Wonder Alexandre Luz Alves, Pedro Henrique Bugatti, Priscila Tiemi Maeda Saito, Douglas Silva Domingues and Alexandre Rossi Paschoal

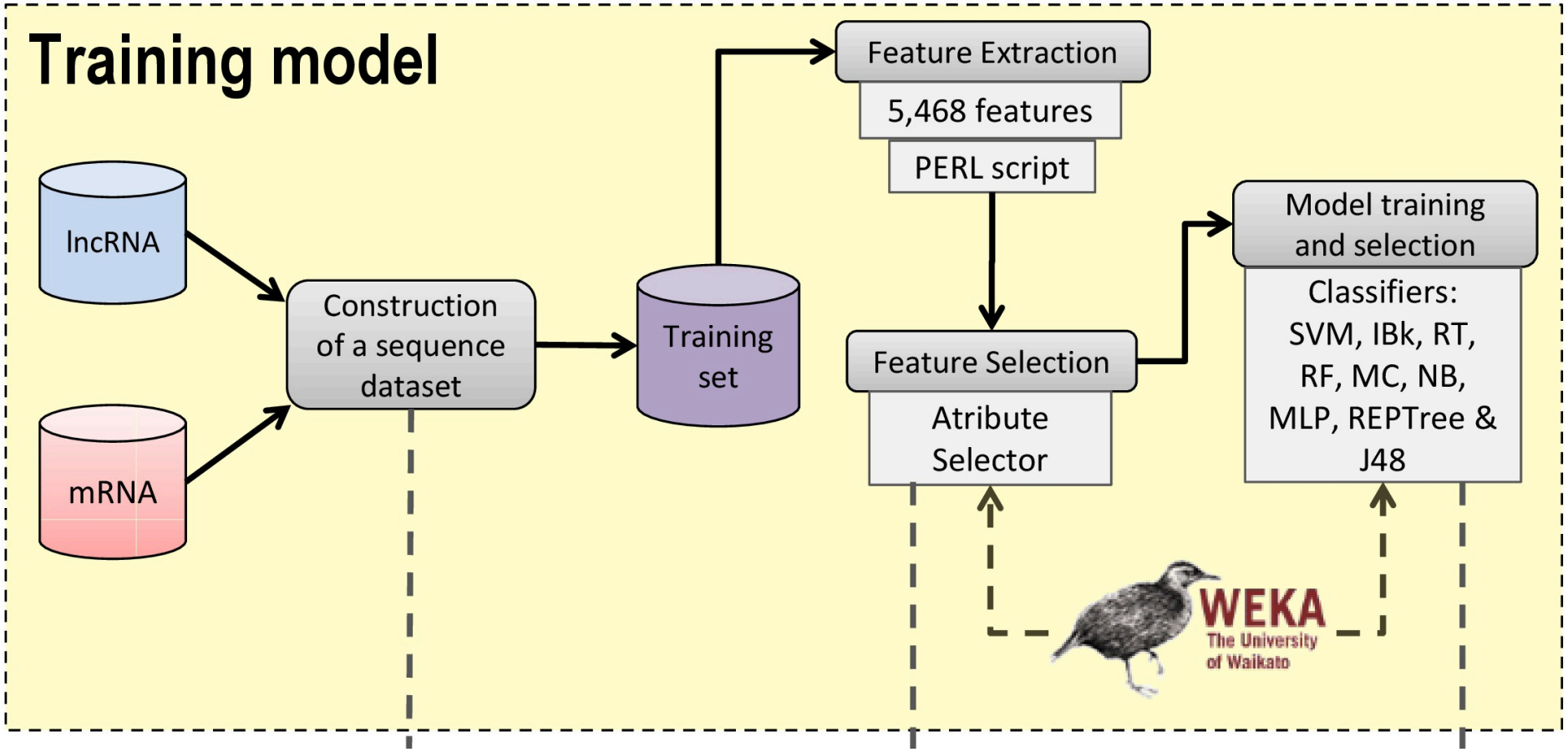
https://doi.org/10.1093/bib/bby034

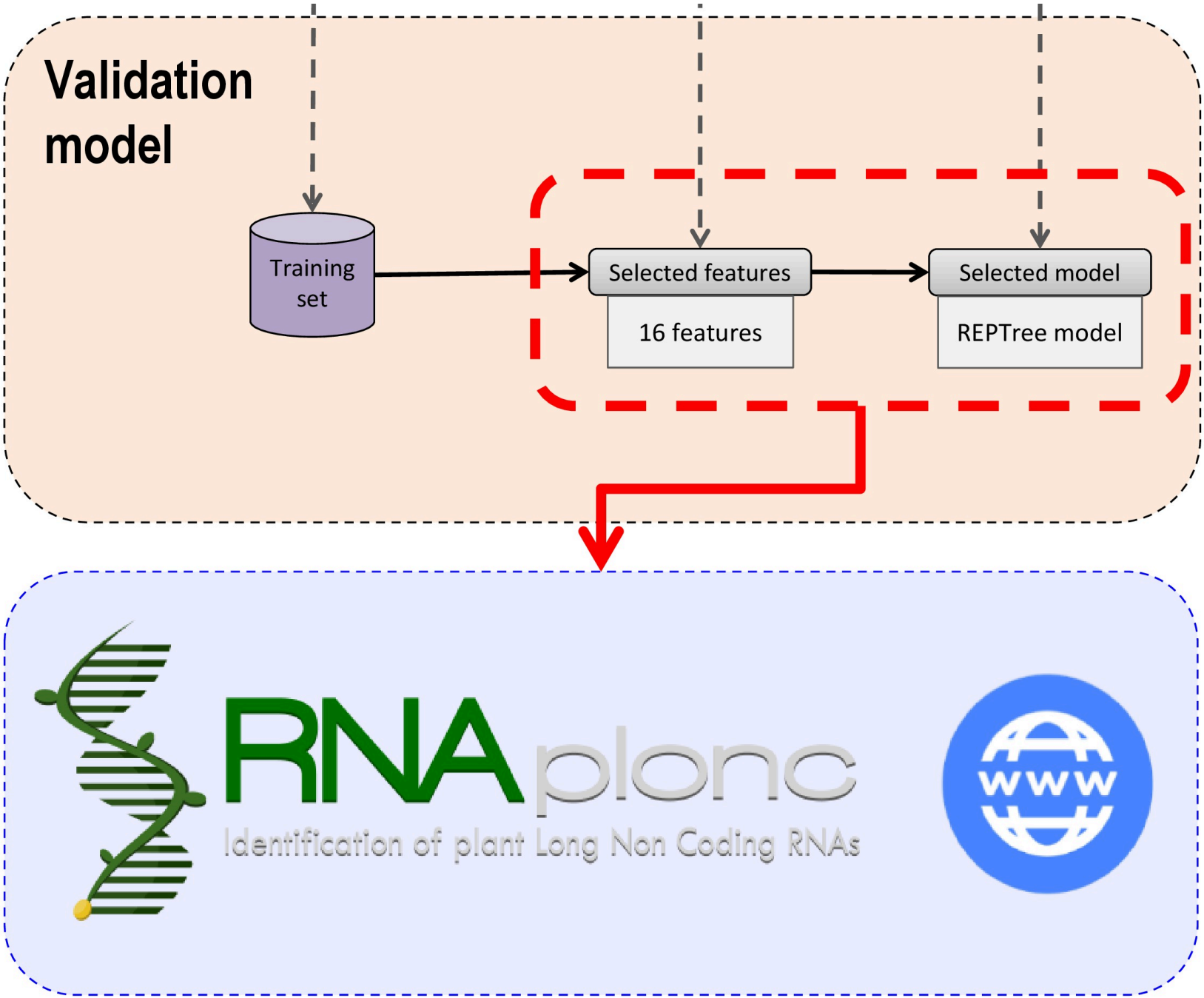
We used 5,468 features




Training model
lncRNA
mRNA
Construction of a sequence dataset
Training set
Feature Extraction
5,468 features
PERL script
Feature Selection
Atribute Selector
Model training and selection
Classifiers: SVM, IBk, RT, RF, MC, NB, MLP, REPTree & J48
WEKA
The University of Waikato
Validation model
Training set
Selected features
16 features
Selected model
REPTree model
RNAplonc
Identification of plant Long Non Coding RNAs
www


Training model
lncRNA
mRNA
Construction of a sequence dataset
Training set


Training model
lncRNA
mRNA
Construction of a sequence dataset
Training set
Feature Extraction
5,468 features
PERL script
Feature Selection
Atribute Selector
Model training and selection
Classifiers: SVM, IBk, RT, RF, MC, NB, MLP, REPTree & J48
WEKA
The University of Waikato


Validation model
Training set
Selected features
16 features
Selected model
REPTree model
RNAplonc
Identification of plant Long Non Coding RNAs
www

16 Best features

Features
GC content
AACG
CCGT
CGCA
CGCT
CGGG
CGTA
TACC
TACG
TCCG
TCGC
Sequence length
score
cdsStop
cdsSizes
cdsPercent

In summary

- A supervised machine learning tool for lncRNA identification

- We got the best 16 features for that in plants

Next Question?

- What the meaning of that?
- What are the contributions of these features for the lncRNA classification in plants?

Features
GC content
AACG
CCGT
CGCA
CGCT
CGGG
CGTA
TACC
TACG
TCCG
TCGC
Sequence length
score
cdsStop
cdsSizes
cdsPercent

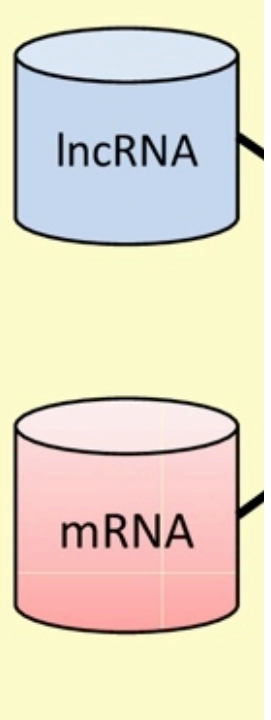
Danilo/Alexandre/Wonder

- Investigation:
  - Feature Selection Methods on RNAplonc tool

- Master student: Robson

One main point is: DATASET

- IS the long (ncRNA) ACTUALLY long?
  - Thinking about datasets to use!
  - MAINLY: negative dataset



Computational Evolutionary Algorithms

- Feature selection methdos:
  - Genetic Algorithm
  - Evolutionary Algorithm
  - Bat Algorithm
  - Artificial Bee Colony
  - Greedy Search
  - Ant Colony Optimization
  - Particle Swarm Optimization

Heuristics feature selection output

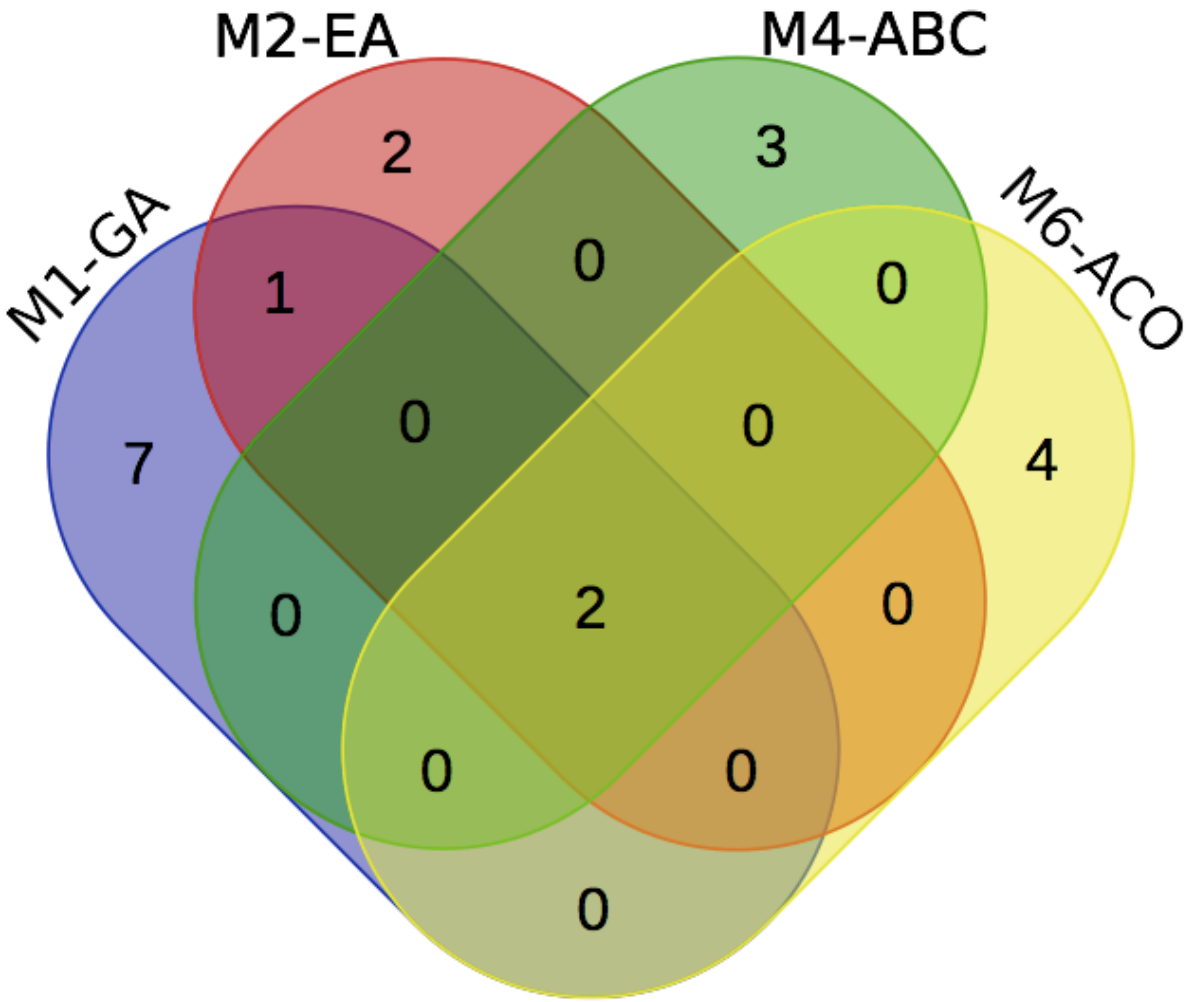
Table 5. Selected Features

M1-GA	M2-EA	M3-BA	M4-ABC	M5-GS	M6-ACO	M7-PSO
ATCCCC	CCGGCA	GCACCC	AGCGGA	AGCACT, CGGAAA, TATGCC, TGACCA, GACTGG	AGCACT	CGCGGA
CCGGCA	GACTAG	GCGTTG	GGGCTA	CGATTC, CATGCC, CGTCAA, CGGAAG, CCTAAA	CCGGGG	CTCGAC
CGCCTC	GAGGGC	TCTGCG	GTCGTC	CCAGTG, AACGCC, GTTCCC, GAATAC, GCTTGC	GAGCCC	GCACGC
CGGAGT	score	score	score	CTCCTA, GATCCC, GTGGCC, CAACCG, CATGCG	GTCGTA	GGGGGG
CGTTAG	cdsSizes	cdsSizes	cdsSizes	CCGTTG, CCTGGA, GGCCTG, GCTGTA, GCCTCC	score	TGACGG
CTAGGT				GCAGCC, AATGCG, CGAGTT, CGGAAT, CAAACG	cdsSizes	score
GGGGGG				GAGTAC, CATCGG, CGCAAG, CCTTAC, GTACCA		cdsSizes
TGACGG				AGCACG, TGCACT, TGGACC, CTGCGG, CGCAAT		
score				AATGCC, AAGGCC, TCTCCG, AAACCG, CTTTAC		
cdsSizes				ACAGTG, GCAGGG, CTCCTG, CCCATG, score, cdsSizes		

Heuristics feature selection output

Table 5. Selected Fe

M1-GA	M2-
ATCCCC	CCGC
CCGGCA	GACT
CGCCTC	GAGC
CGGAGT	sco
CGTTAG	cdsSi
CTAGGT	
GGGGGG	
TGACGG	
score	
cdsSizes	



CO	M7-PSO
ACT	CGCGGA
GGG	CTCGAC
CCC	GCACGC
GTA	GGGGGG
re	TGACGG
izes	score
	cdsSizes

Conclusion

- Two ORF features (CDS size and score) are enough to lncRNA identification !?
  - Crazy?!?!?!?
- … Let's try another idea

Conclusion

- Two ORF features (CDS size and score) are enough to lncRNA identification !?

- … Let's try another idea
- New Test: REMOVE this two ORF features

We found another ORF features

Table 9. New Selected Features

M1-GA	M4-ABC	M6-ACO
CCAGG, TCTGC	GGGTCG	CCGGA, CCTGG, GTTGC
CATCAA, CTCATG	cdsStop	TGCGG, AAGGCC, ACCTCC
CTGCAG, GAAGGA	cdsPercent	ACGGAG, ACTGGG, AGAGCT
CCGGGG, GGAATT		AGCTGG, ATCTGG, CAAGGA
GGACCC		CAGAGT, CTTGAC, GACAGC
cdsStop		GAGGGG GGGTGC, GGTTAT
cdsPercent		TGCTGC, TGGGCT, GCTGTT
		GCTCTG, GCCTTC, GATGAG
		TTCTGG, cdsStop, cdsPercent

Final conclusion

- Is the long ncRNA actually long?
- What is there available of lncRNA public data?

THANK YOU

ANY

Question?

Google it

CAPES

Bioinformática UTFPR

UTFPR

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ