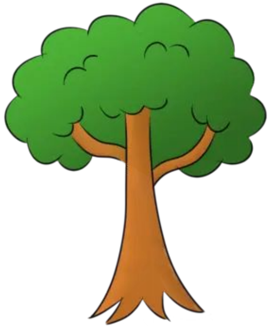
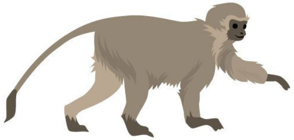
Codon Frequency Classification Project

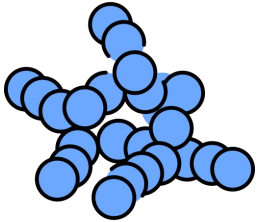
Sprint 3

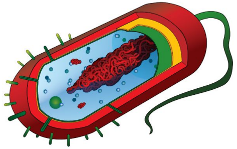
Pier Bruno Pompilii

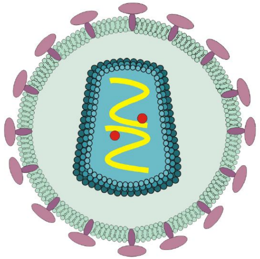
Can you identify this species by eye?



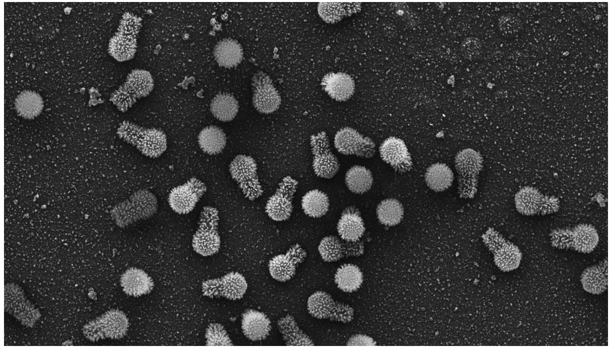








Now?



https://www.pourlascience.fr/sd/microbiologie/des-virus-geants-tres-bacteriens-13161.php

https://sec-sem.blogspot.com/2011/01/bacteria-cultivation-mini-sem-image.html

We can use DNA to differentiate, but ….

Can the usage of different codons be used to classify species by Kingdoms?

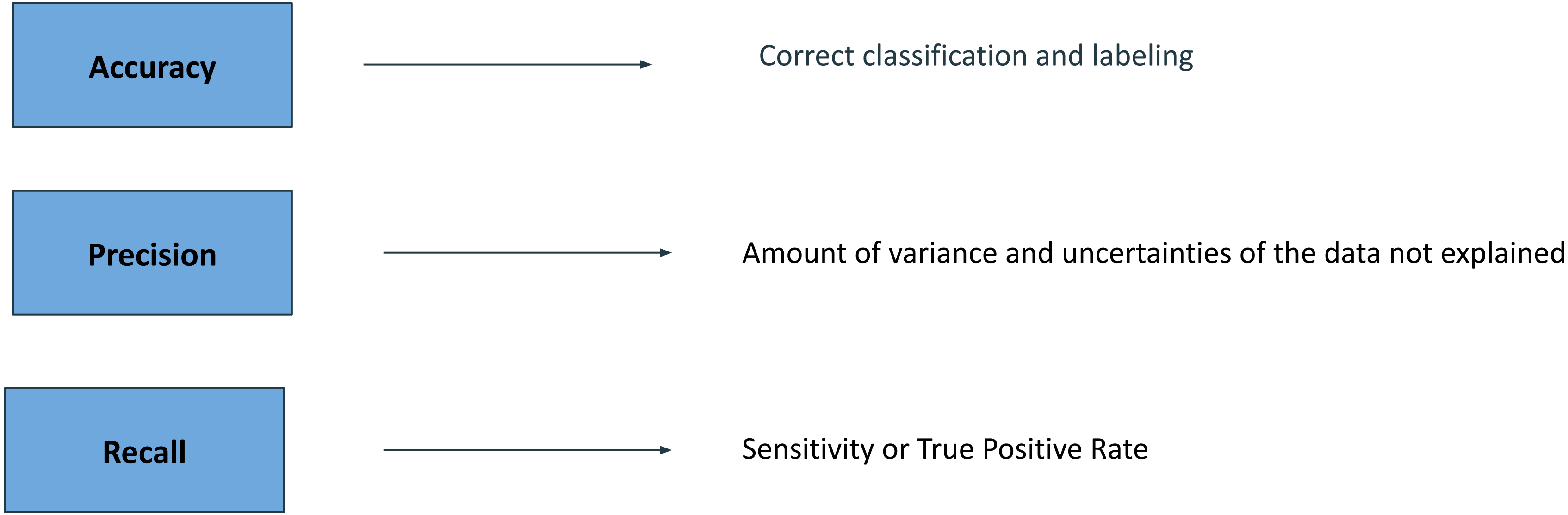
12 K Species	Phylogenetic information
64 Codons	Usage frequencies.
5 Kingdoms	Animalia, Plantae, Bacteria, Archaea, Virus

Objective

Attempt to classify codon usage in terms of lineage, by using machine learning methods to identify this genomics and evolutionary differences

Approach

Statistical Metrics for the ML models



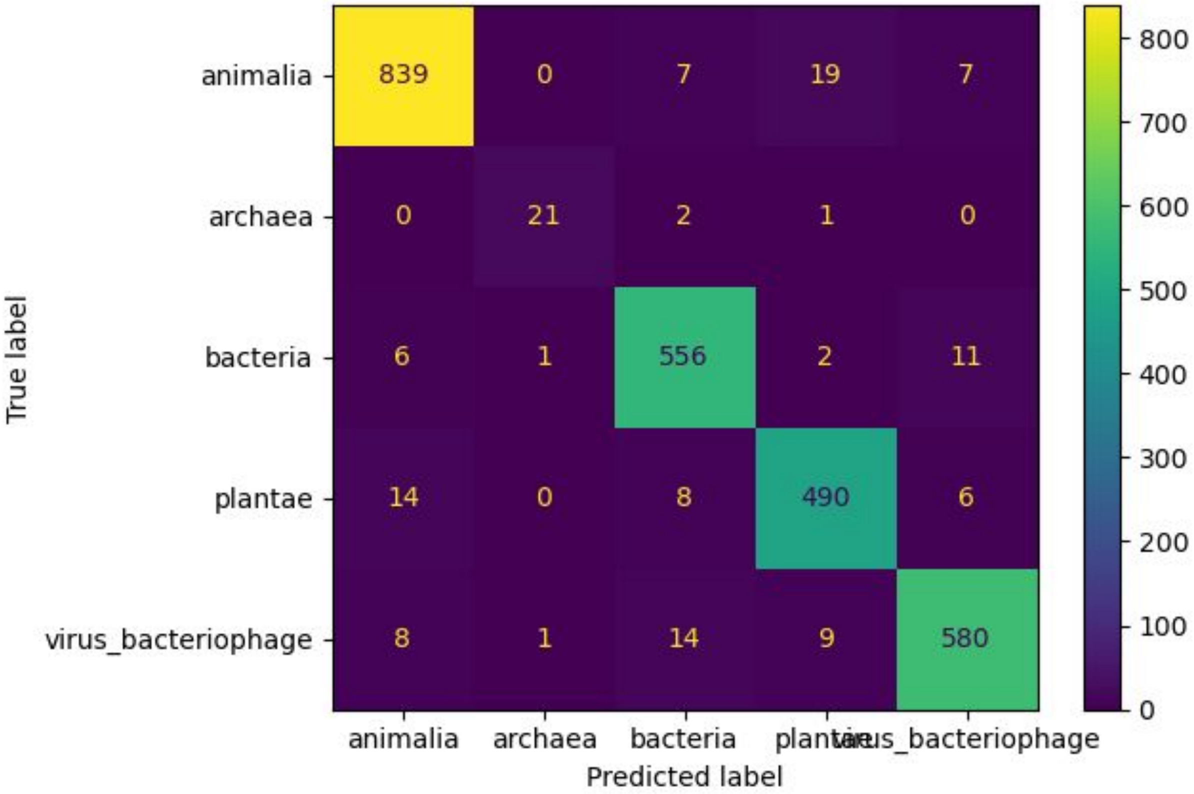
Accuracy → Correct classification and labeling

Precision → Amount of variance and uncertainties of the data not explained

Recall → Sensitivity or True Positive Rate

Which model ?

KNN



True label \ Predicted label	animalia	archaea	bacteria	plantae	virus_bacteriophage
animalia	839	0	7	19	7
archaea	0	21	2	1	0
bacteria	6	1	556	2	11
plantae	14	0	8	490	6
virus_bacteriophage	8	1	14	9	580

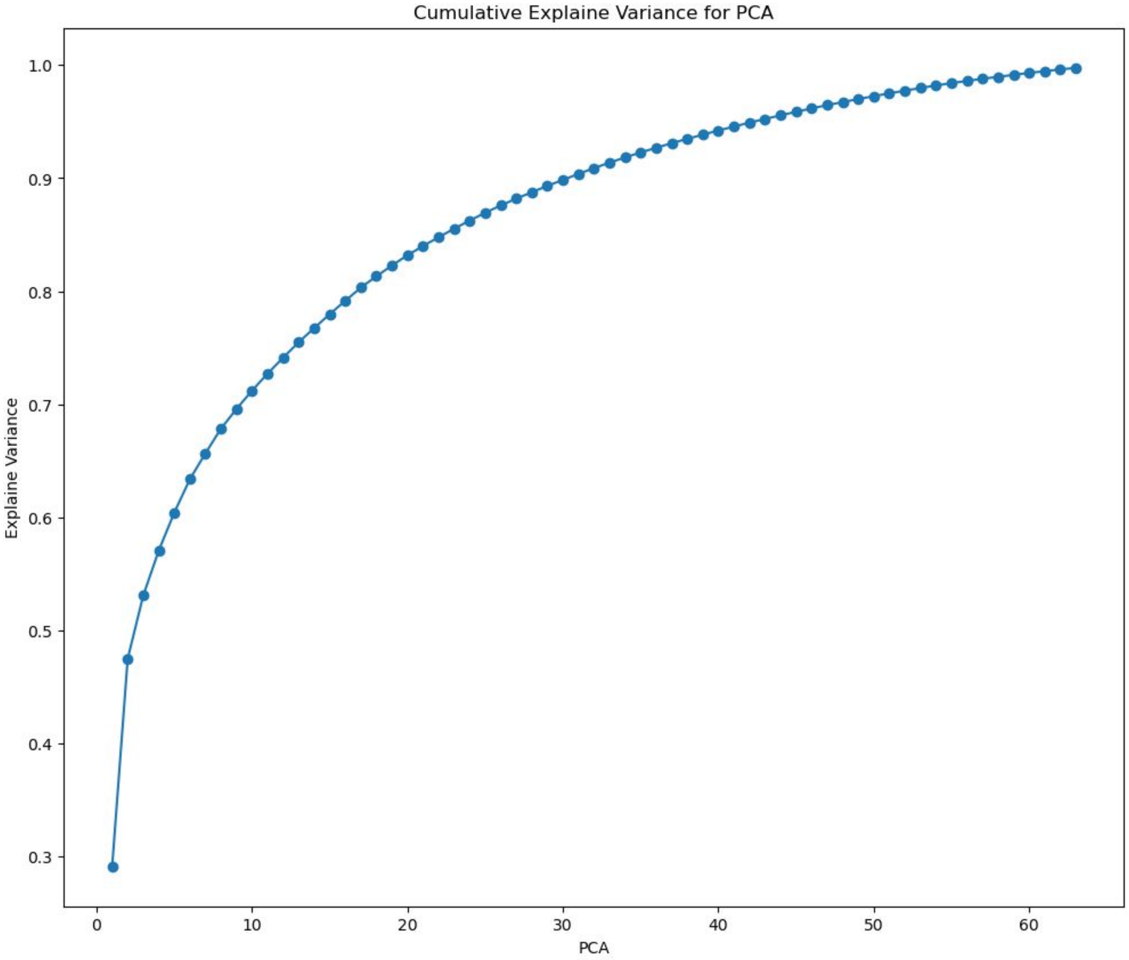
Accuracy = 96%

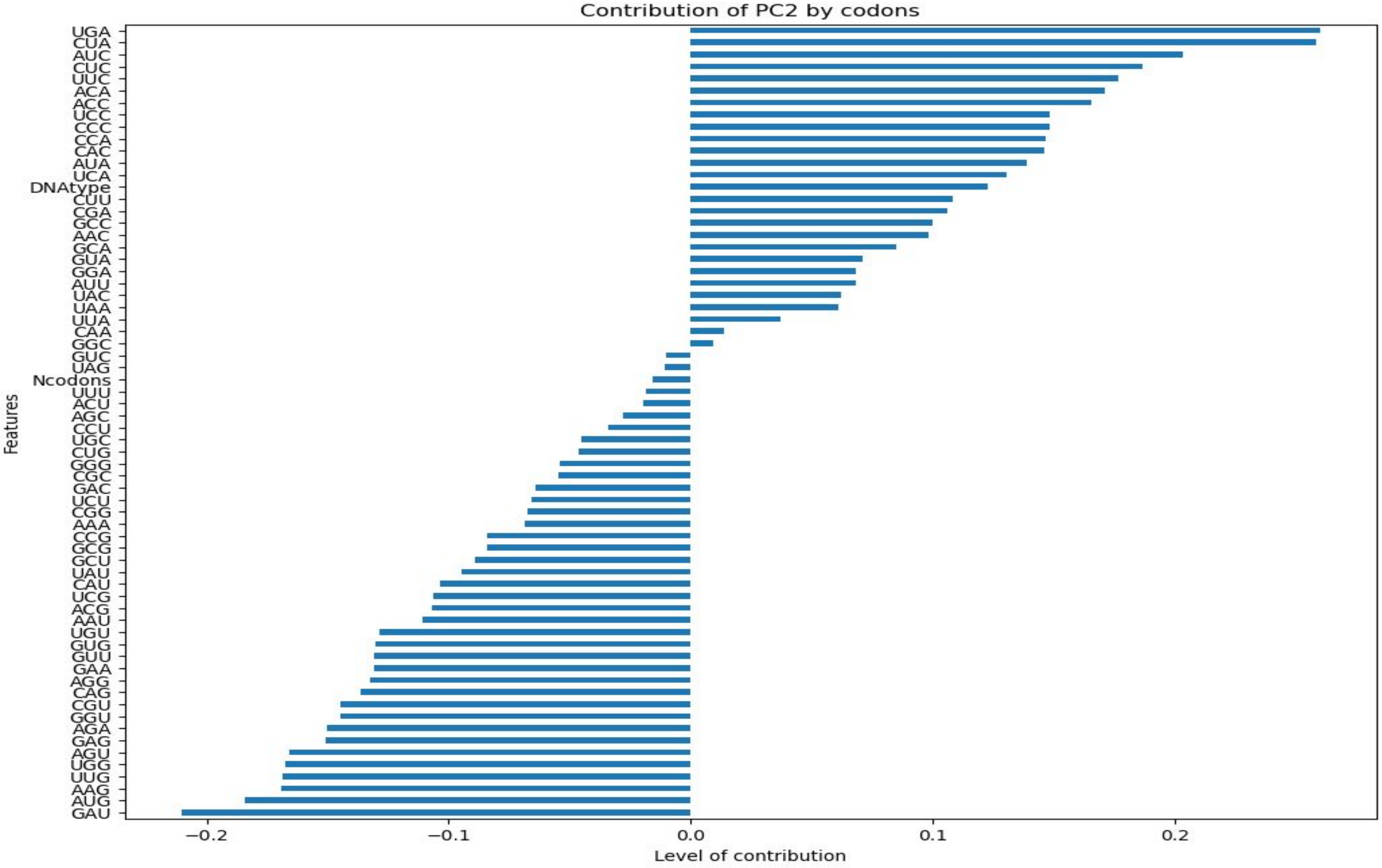
Precision = 95%

Recall = 94%

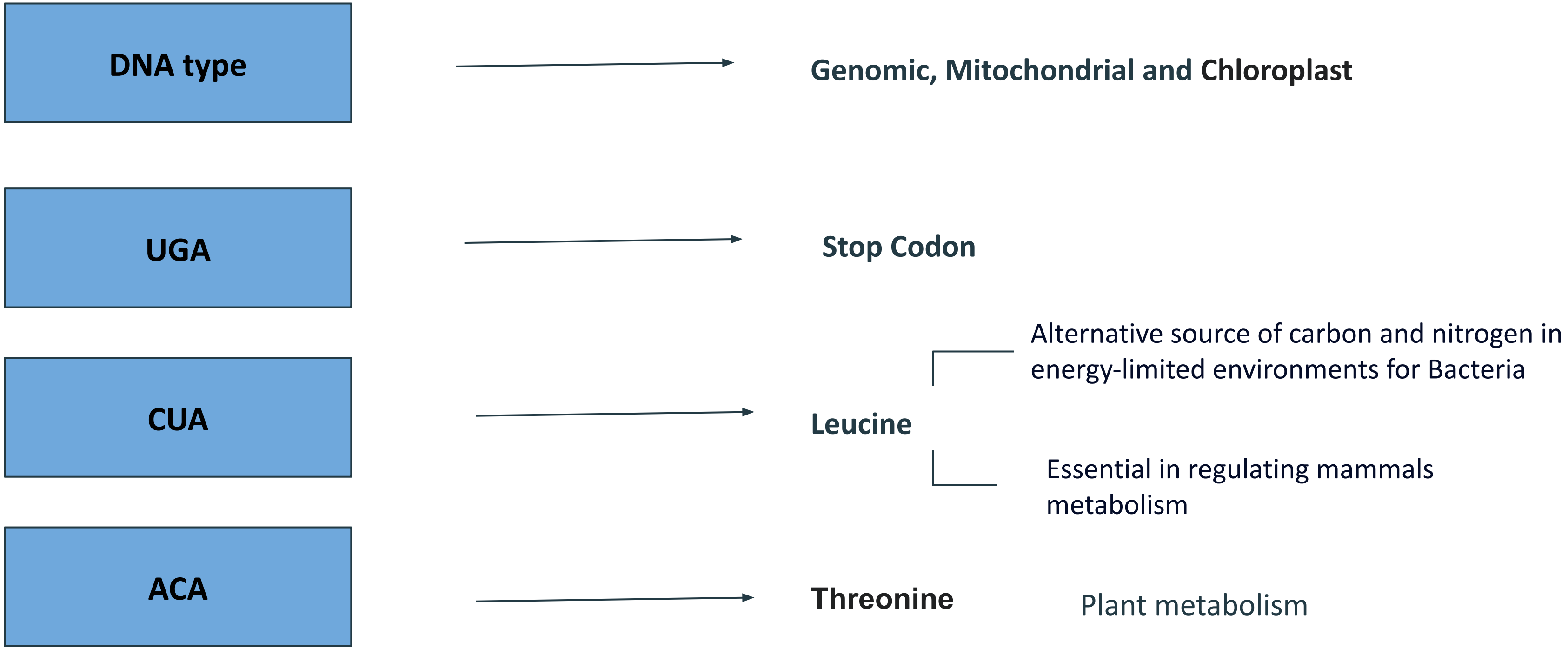
PCA Analysis

The principal components contributing to the classification task are **PC1, PC2** and **PC3**




Contribution of PC2 by codons
Features
UGA
CUA
AUC
CUC
UUC
ACA
ACC
UCC
CCC
CCA
CAC
AUA
UCA
DNAtype
CUU
CGA
GCC
AAC
GCA
GUA
GGA
AUU
UAC
UAA
UUA
CAA
GGC
GUC
UAG
Ncodons
UUU
ACU
AGC
CCU
UGC
CUG
GGG
CGC
GAC
UCU
CGG
AAA
CCG
GCG
GCU
UAU
CAU
UCG
ACG
AAU
UGU
GUG
GUU
GAA
AGG
CAG
CGU
GGU
AGA
GAG
AGU
UGG
UUG
AAG
AUG
GAU
−0.2
−0.1
0.0
0.1
0.2
Level of contribution

Some Contributors



DNA type → **Genomic, Mitochondrial and Chloroplast**

UGA → **Stop Codon**

CUA → **Leucine**

- Alternative source of carbon and nitrogen in energy-limited environments for Bacteria
- Essential in regulating mammals metabolism

ACA → **Threonine**

Plant metabolism

Impact

Improving Taxonomic Classification

Refining kingdom definitions or to be used for the discovery of new species

Genetic Research

Facilitating the discovery of genetic markers that can be used for species identification

Bioinformatics

Machine Learning can answer biological questions for research

Thanks!