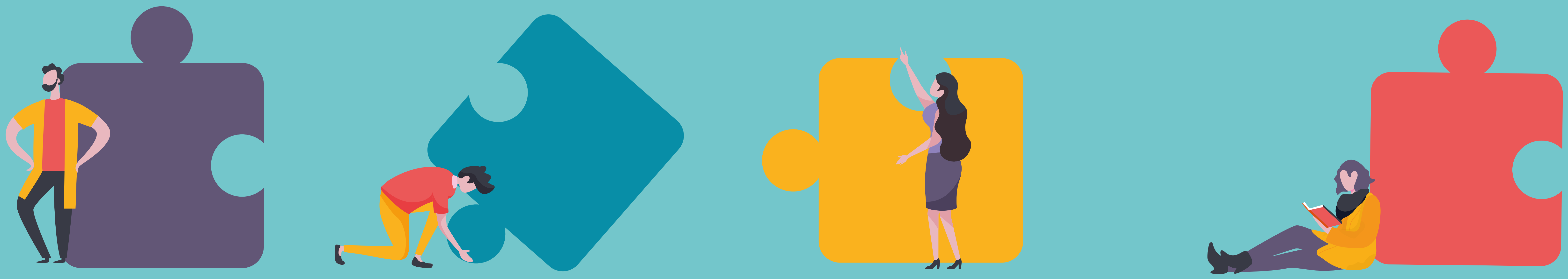
Reconstruct the Information

Presented by :

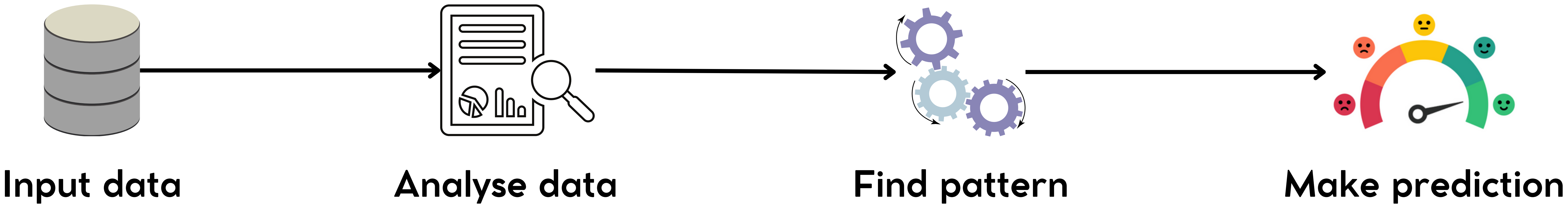
- Houria BRAIKIA
- Tahar
- Theo
- Nicolas

Houria



Machine Learning

ML is increasingly recognized for its capacity to accelerate research across a range of scientific disciplines by efficiently identifying patterns within vast and complex datasets.



In biology, especially in **microbiome studies**, various experimental techniques generate vast datasets, which ML analyzes for insights and predictions, playing a crucial role in advancing scientific understanding.

Microbiome Data

Microbiome data refers to the collective genetic material of microorganisms inhabiting a specific environment, such as the human gut, soil, or water.

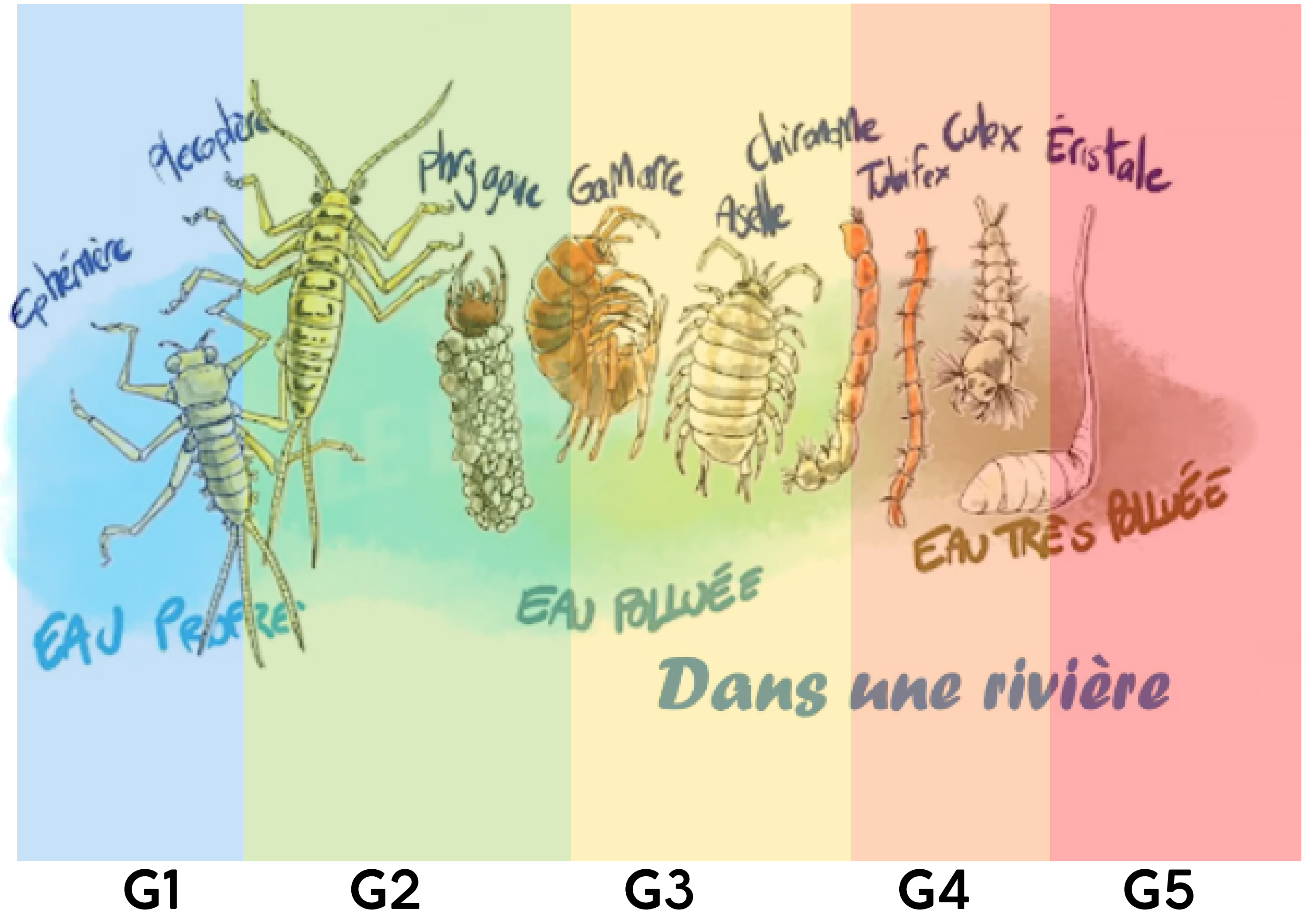
- **Count data:** Indicates the number of occurrences of each species in a sample.
- **Presence/absence data:** Simply indicate whether each species is present or absent in a sample.
- **DNA sequence data:** Provide the DNA or RNA sequences.

Biomonitoring

The availability of microbiome data from marine environments has facilitated the assessment of environmental health.

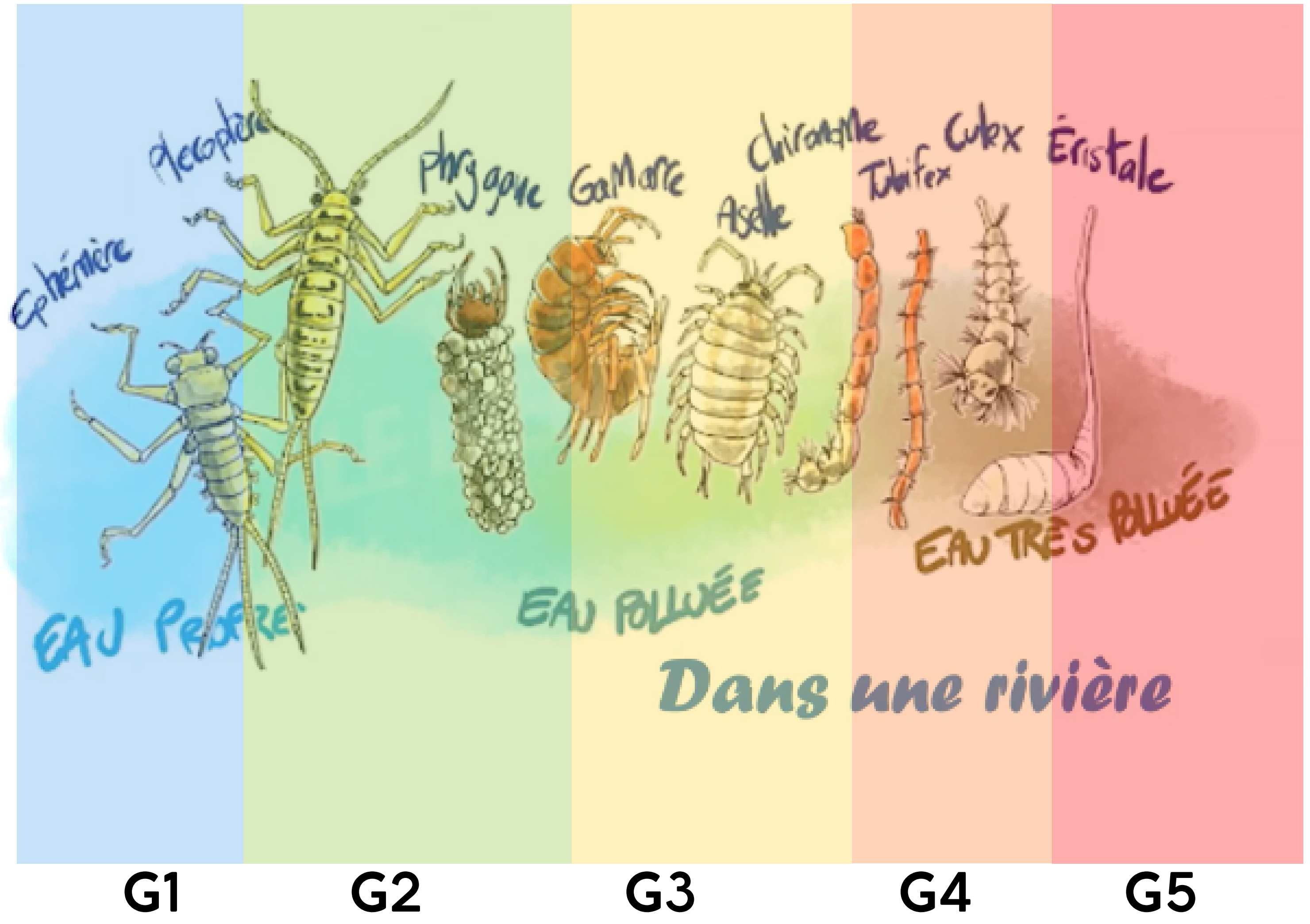
Ecological Quality

The availability of microbiome data from marine environments has facilitated the assessment of environmental health.



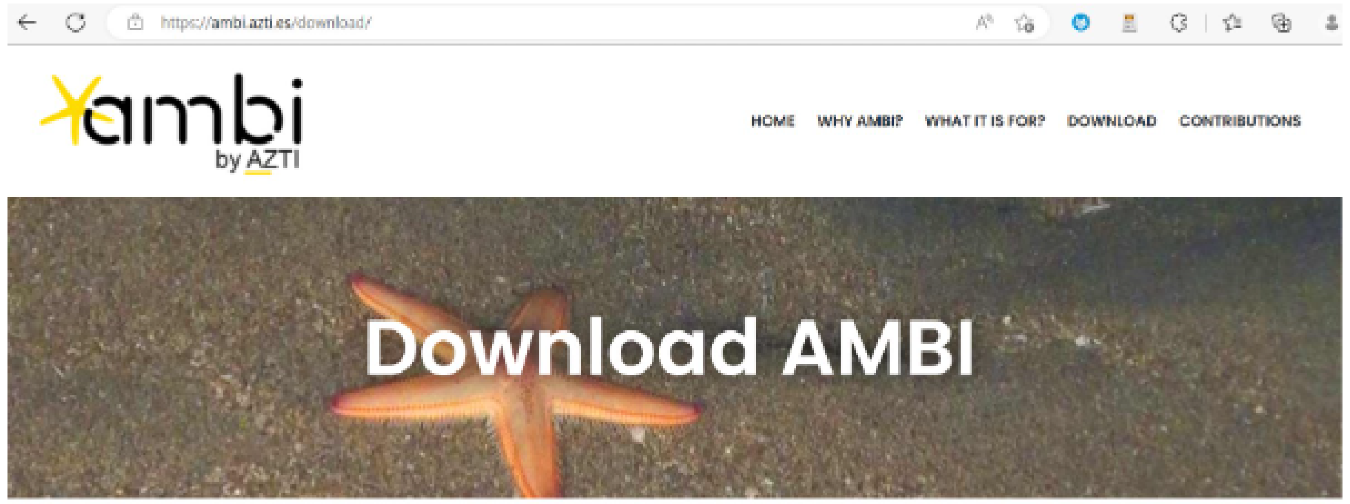
Ecological Quality

The availability of microbiome data from marine environments has facilitated the assessment of environmental health.



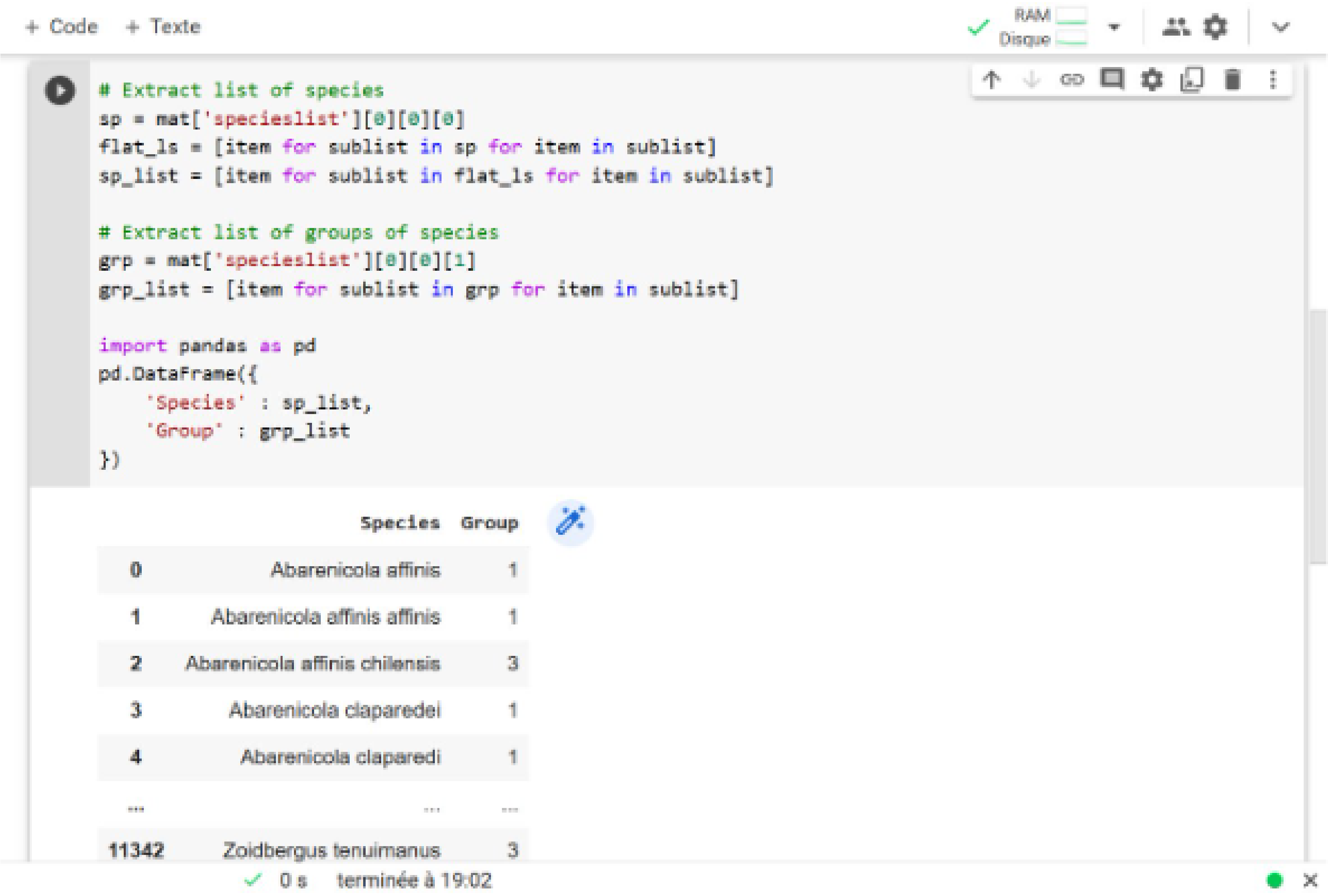
AMBI = ((0 x %G1) + (1.5 x %G2) + (3 x %G3) + (4.5 x %G4) + (6 x %G5)) / 100

Ecological Quality



The new AMBI 6.0 version is out, with new functionalities!

https://ambi.azti.es/download




```python
# Extract list of species
sp = mat['specieslist'][0][0][0]
flat_ls = [item for sublist in sp for item in sublist]
sp_list = [item for sublist in flat_ls for item in sublist]

# Extract list of groups of species
grp = mat['specieslist'][0][0][1]
grp_list = [item for sublist in grp for item in sublist]

import pandas as pd
pd.DataFrame({
    'Species' : sp_list,
    'Group' : grp_list
})
```

	Species	Group
0	Abarenicola affinis	1
1	Abarenicola affinis affinis	1
2	Abarenicola affinis chilensis	3
3	Abarenicola claparedei	1
4	Abarenicola claparedi	1
...	...	...
11342	Zoidbergus tenuimanus	3

Ecological Quality

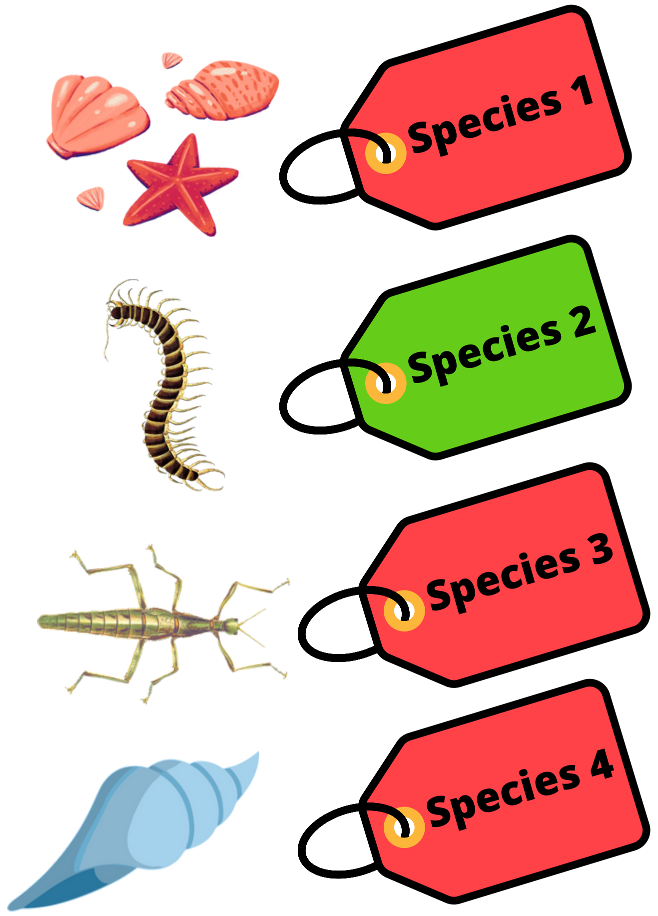
$$AMBI = ((0 \times \%G1) + (1.5 \times \%G2) + (3 \times \%G3) + (4.5 \times \%G4) + (6 \times \%G5)) / 100$$

AMBI values	Ecological quality class
< 1.2	Very good
Between 1.2 and 3.3	Good
Between 3.3 and 4.3	Moderate
Between 4.3 and 5.5	Bad
$>=5.5$	Very bad

Two fundamental pieces of information are needed to calculate the biotic index:

species group and abundance.

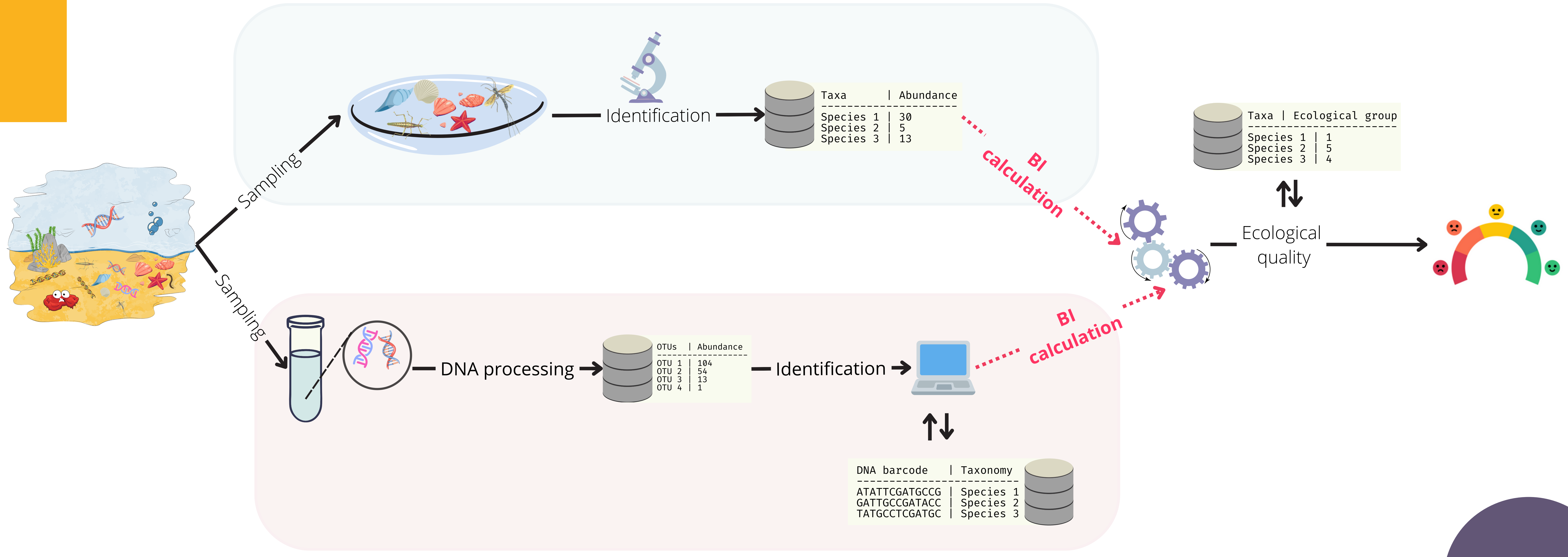
Ecological Quality



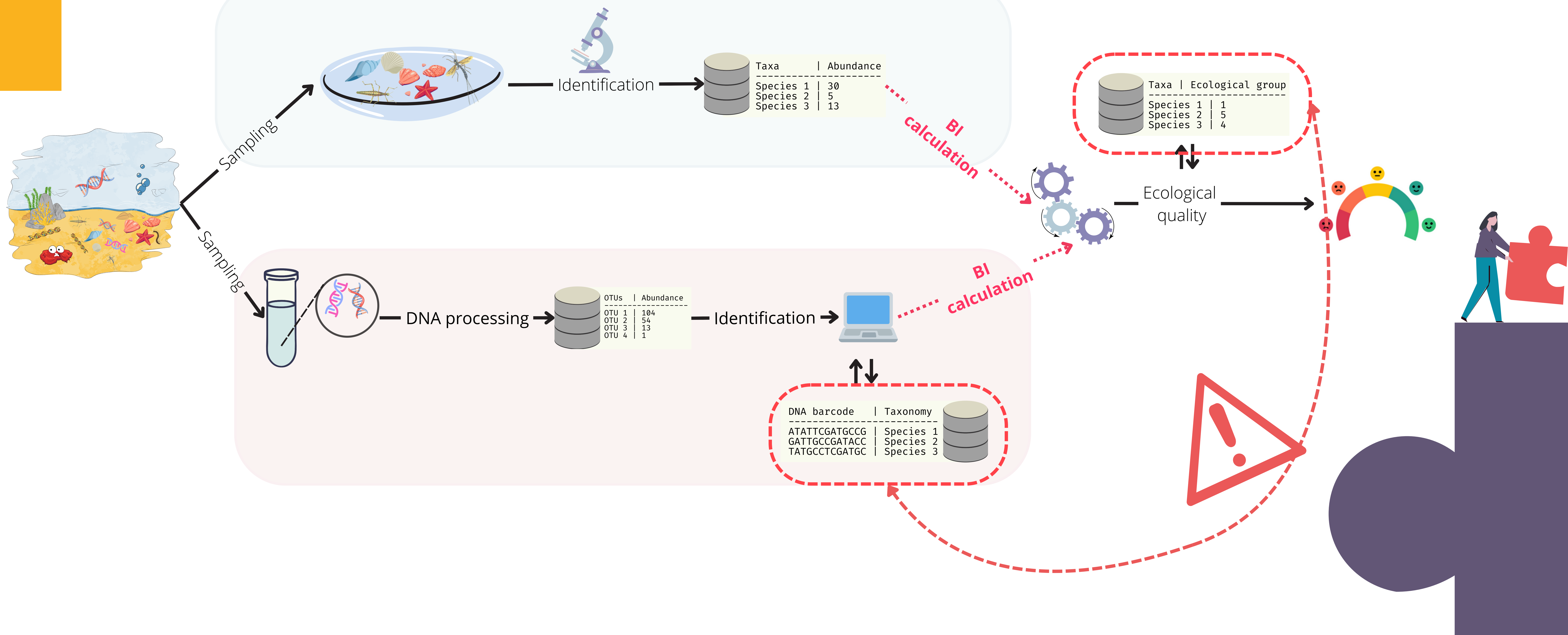
Many species belong to group 5

A polluted environment

Ecological Quality Assesment



Incomplete Information



Reconstruct the Information

Supervised Machine Learning (SML) approache have been proposed to generate predictive models of BI values from eDNA data, even with unassigned taxa.

	OTU_ID	Total_Abund_Otu	FHF1-St1-Gr1-A	FHF1-St1-Gr1-B	FHF1-St1-Gr1-C	FHF1-St1-Gr2-A	FHF1-St2-Gr1-C	FHF1-St2-Gr2-A	FHF1-St2-Gr2-B	FHF1-St3-Gr1-A	...	FHF5-St4-Gr2-B	FHF5-St4-Gr2-C	FHF5-St5-Gr1-A	FHF5-St5-Gr1-B	FHF5-St5-Gr1-C	FHF5-St5-Gr2-A	FHF5-St5-Gr2-B	FHF5-St5-Gr2-C
0	OTU0	662523	57	2	6	226	14782	18018	9607	52877	...	3562	199	134	37	58	342	125	795
1	OTU1	574100	110	18	3124	84	12	10	26	60	...	764	132	4930	2524	6925	179	91	574
2	OTU2	698457	4846	38	4138	6178	2900	5306	11986	4026	...	26016	51776	5043	3341	1879	3323	1921	15264
3	OTU3	532877	7043	39	7485	26166	35325	26655	14690	3930	...	5	113	353	1675	537	2756	3764	7154
4	OTU4	463490	1990	68	2140	3919	2573	3563	13387	1149	...	402	839	957	1202	954	1788	1811	3256
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
12327	OTU12327	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
12328	OTU12328	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
12329	OTU12329	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
12330	OTU12330	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
12331	OTU12331	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

12332 rows × 127 columns

Reconstruct the Information

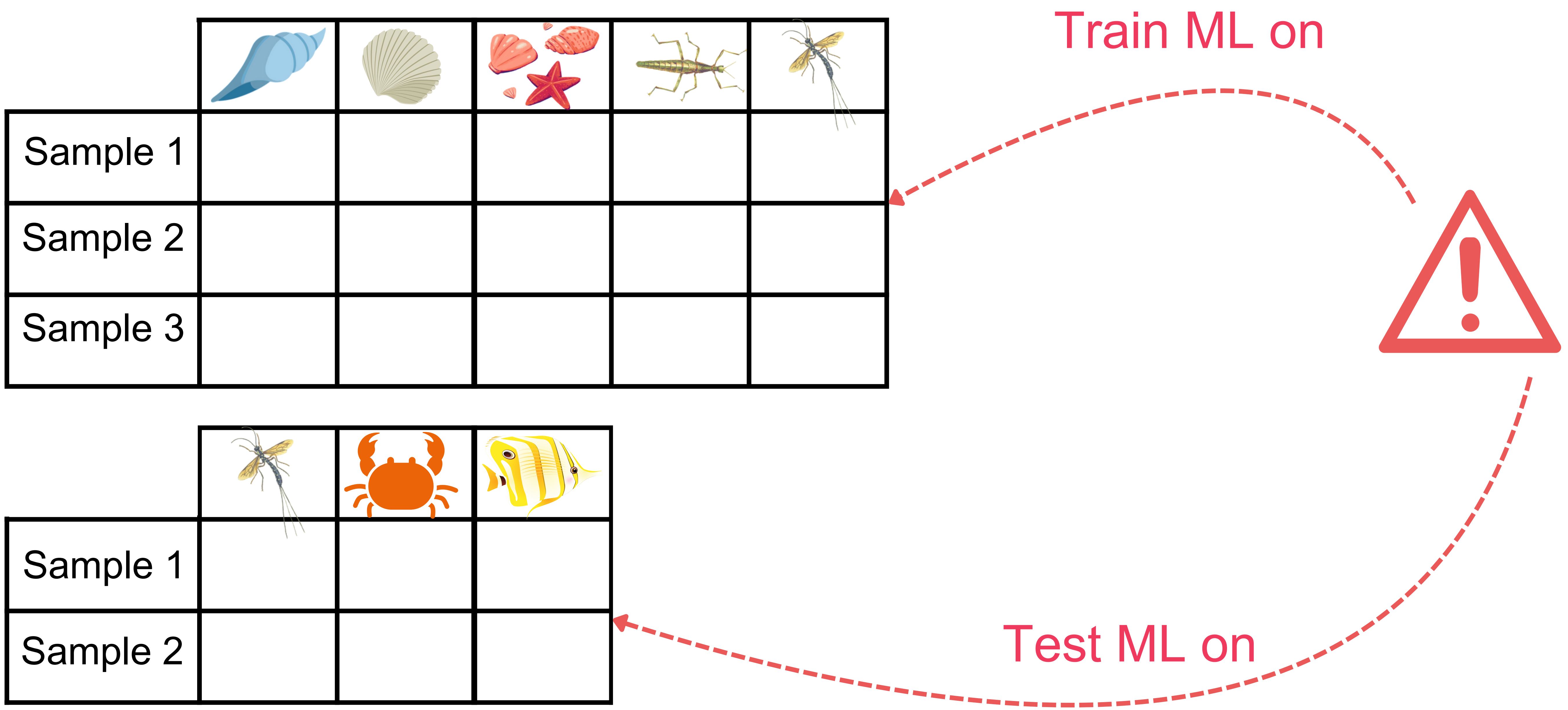
Supervised Machine Learning (SML) approache have been proposed to generate predictive models of BI values from eDNA data, even with unassigned taxa.

Code Blame 153 lines (153 loc) · 9.82 KB

Sample	Farm	Replicate	Distance	AMBI	EQ	Latitude	Longitude
AK_0_G1_RepA	Aukrasanden	1	0	5,96938	5	62°46.935 N	6°55.399 E
AK_0_G1_RepB	Aukrasanden	2	0	5,96938	5	62°46.935 N	6°55.399 E
AK_0_G1_RepC	Aukrasanden	3	0	5,96938	5	62°46.935 N	6°55.399 E
AK_0_G2_RepA	Aukrasanden	4	0	5,96853	5	62°46.935 N	6°55.399 E
AK_0_G2_RepB	Aukrasanden	5	0	5,96853	5	62°46.935 N	6°55.399 E
AK_0_G2_RepC	Aukrasanden	6	0	5,96853	5	62°46.935 N	6°55.399 E
AK_1530_G1_RepA	Aukrasanden	1	1530	1,30127	2	62°46.457 N	6°56.861 E
AK_1530_G1_RepB	Aukrasanden	2	1530	1,30127	2	62°46.457 N	6°56.861 E
AK_1530_G1_RepC	Aukrasanden	3	1530	1,30127	2	62°46.457 N	6°56.861 E
AK_1530_G2_RepA	Aukrasanden	4	1530	1,38916	2	62°46.457 N	6°56.861 E
AK_1530_G2_RepB	Aukrasanden	5	1530	1,38916	2	62°46.457 N	6°56.861 E
AK_1530_G2_RepC	Aukrasanden	6	1530	1,38916	2	62°46.457 N	6°56.861 E
AK_630_G1_RepA	Aukrasanden	1	630	2,40892	2	62°46.623 N	6°55.701 E
AK_630_G1_RepB	Aukrasanden	2	630	2,40892	2	62°46.623 N	6°55.701 E
AK_630_G1_RepC	Aukrasanden	3	630	2,40892	2	62°46.623 N	6°55.701 E

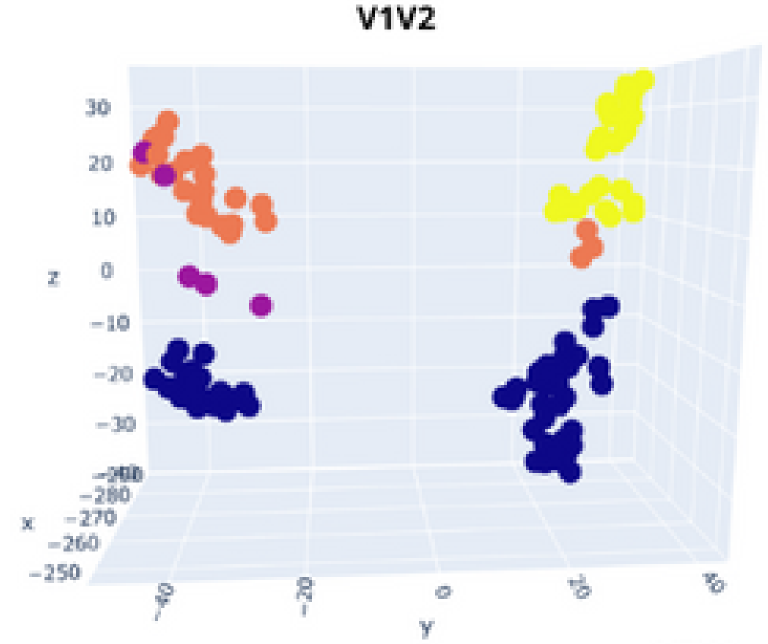
Other Incomplete Information

a key limitation arises when applying SML to samples with different compositional data, necessitating the development of new models using new training data.

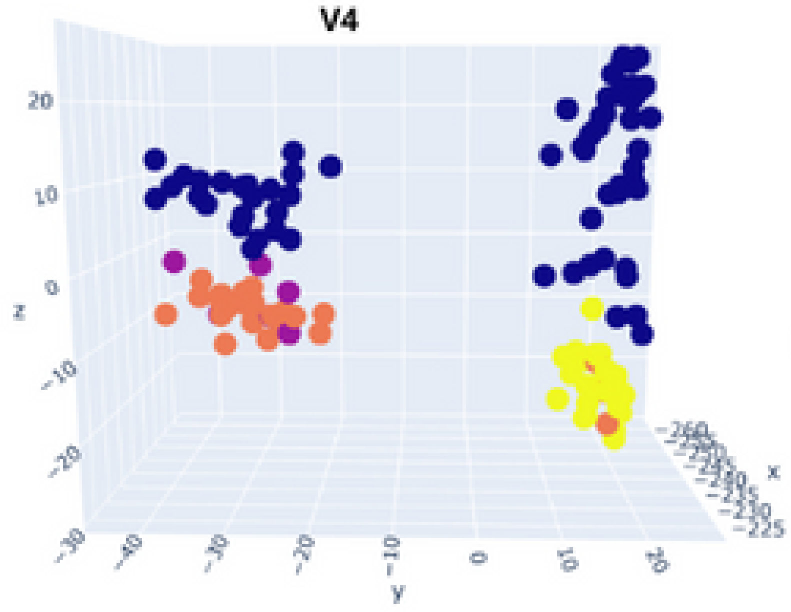


Reconstruct the Information Again

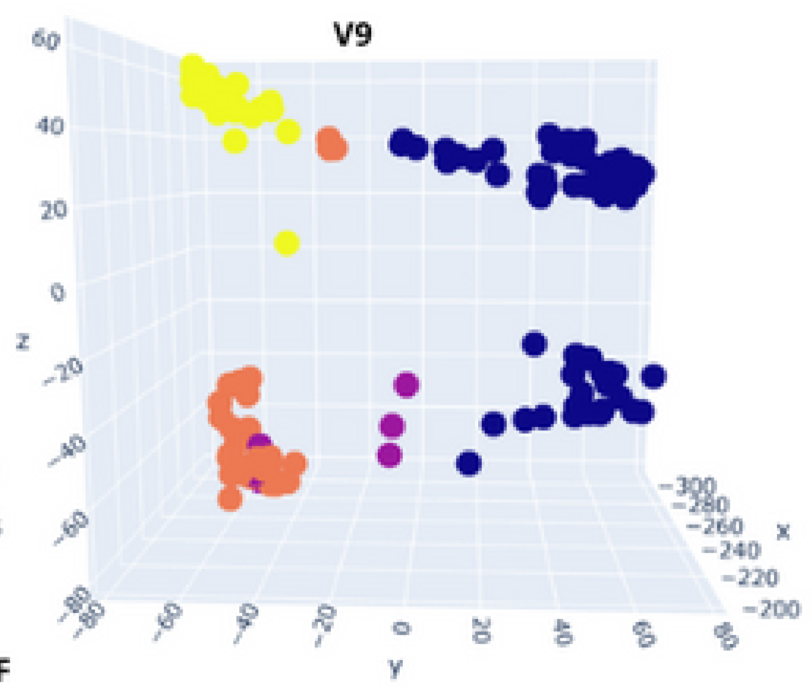
We proposed a solution that involves predicting EQ with minimal reference samples using USML.



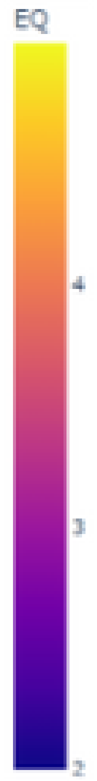
R statistic = 0.593; P-value= 0.001



R statistic = 0.395; P-value= 0.001

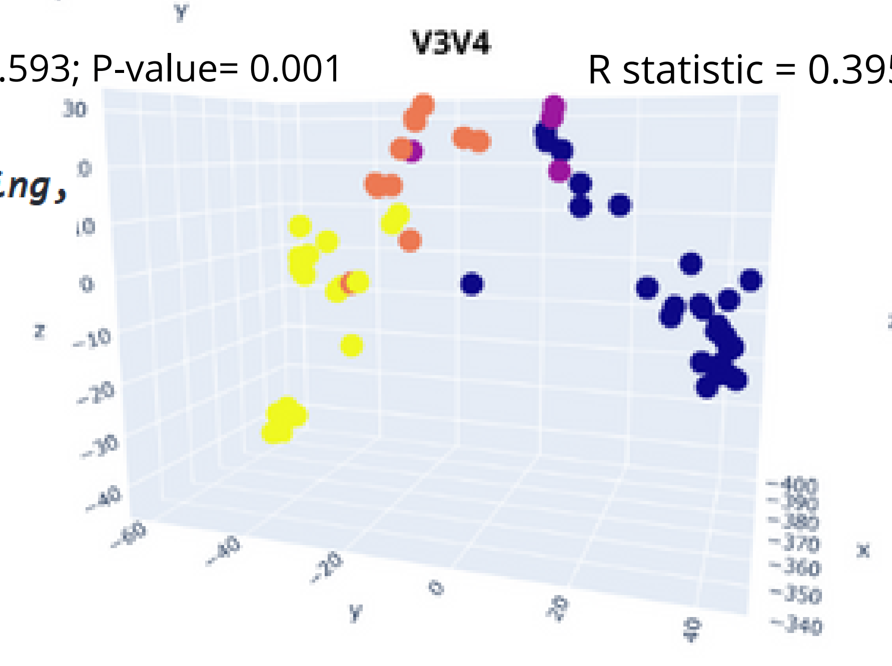


R statistic = 0.894; P-value= 0.001

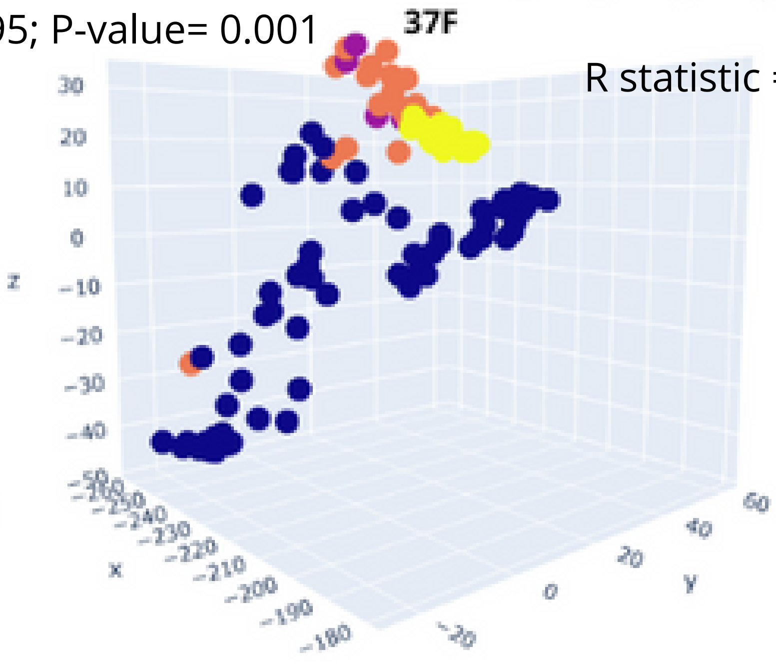


SCIKIT-BIO

```
skbio.stats.distance.anosim(distance_matrix, grouping,
column=None, permutations=999)
```



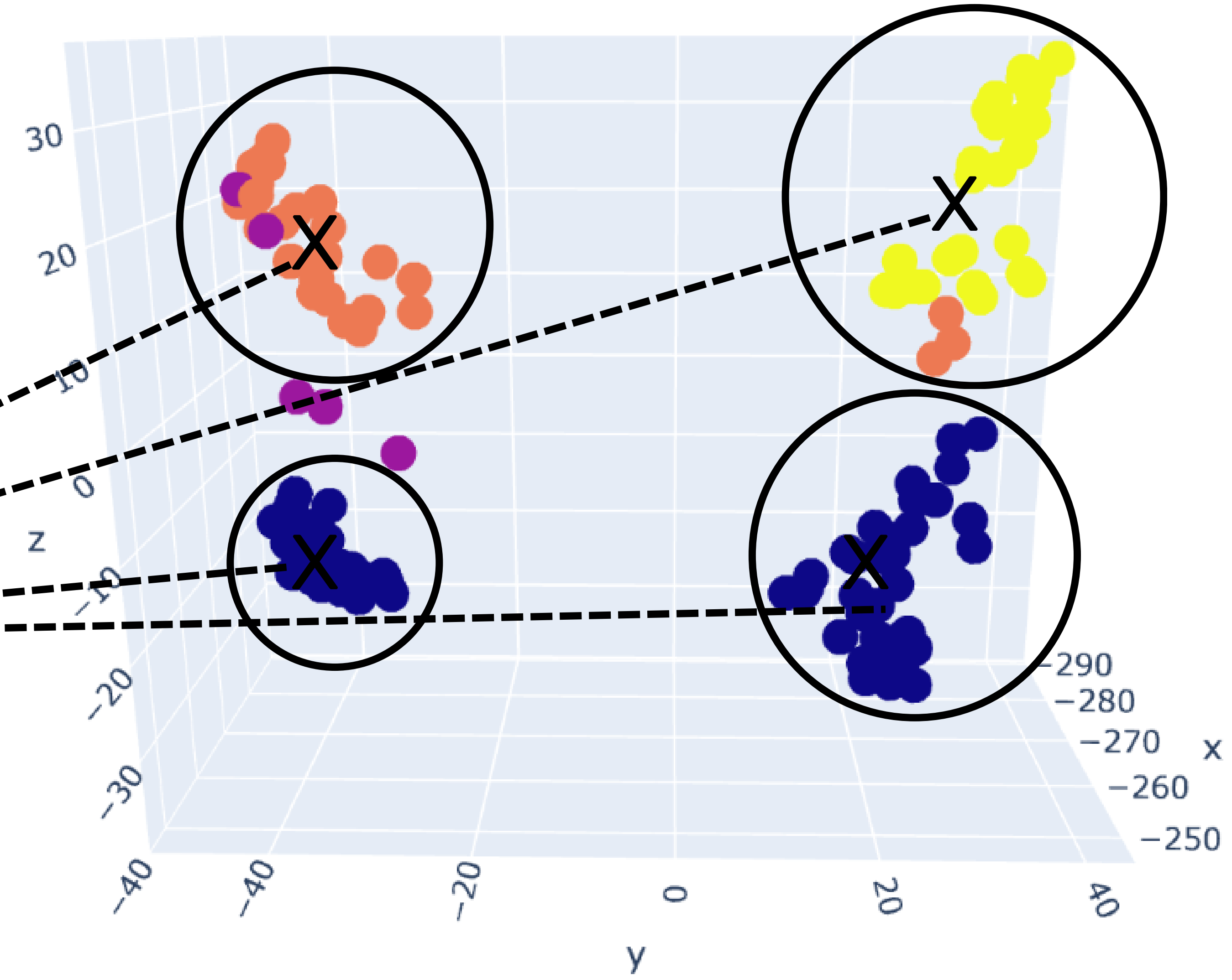
R statistic = 0.794; P-value= 0.001



R statistic = 0.14; P-value= 0.001

Reconstruct the Information Again

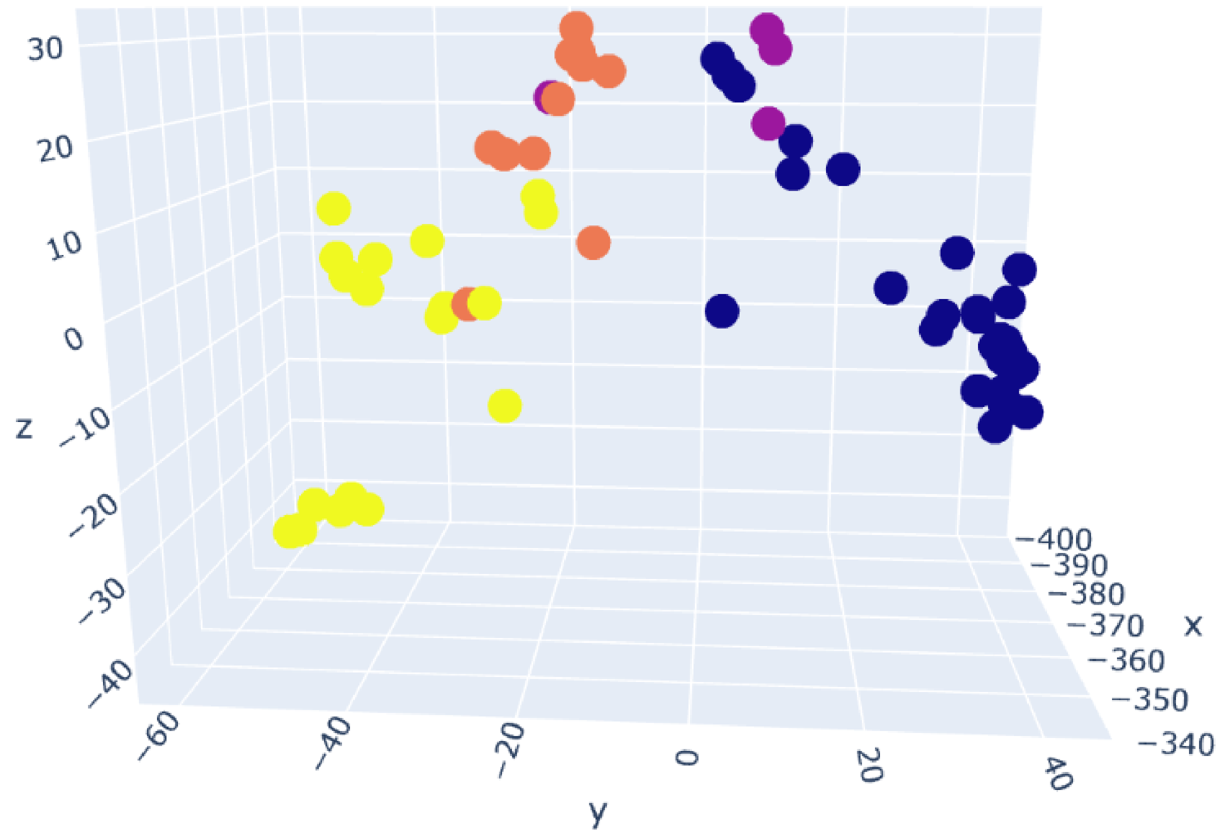
We proposed a solution taht involves predicting EQ with minimal reference samples using USML.

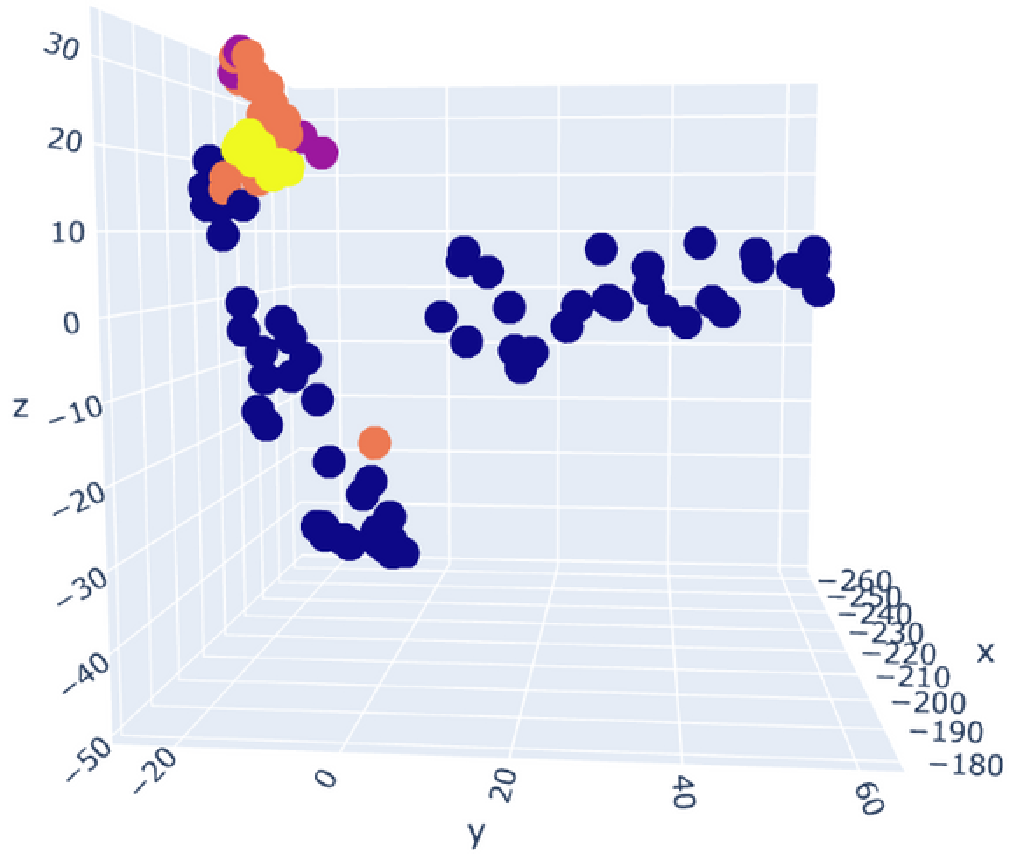


Centroids

Challenges

This approach demonstrates effective predictive capabilities, notably for eukaryotic markers, while highlighting challenges with dispersed bacterial data and single taxonomic group markers like foraminifera.





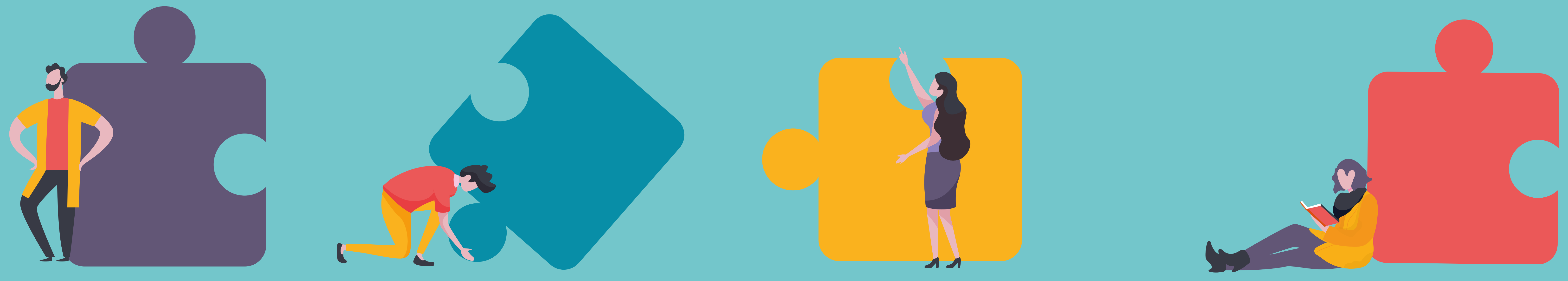
Conclusion

Our approach presents a promising solution to address the persistent challenge of insufficient data in reference databases.

Despite the progress made, it's important to note that the problem remains unsolved, highlighting the need for further research and innovation in this area.

Although supervised machine learning (SML) generally outperforms unsupervised machine learning (USML), the problem of missing labels hinders its use.

Tahar



Reconstructing Hidden Ground Truths via Votes

Condorcet, aka, Father of Crowdsourcing

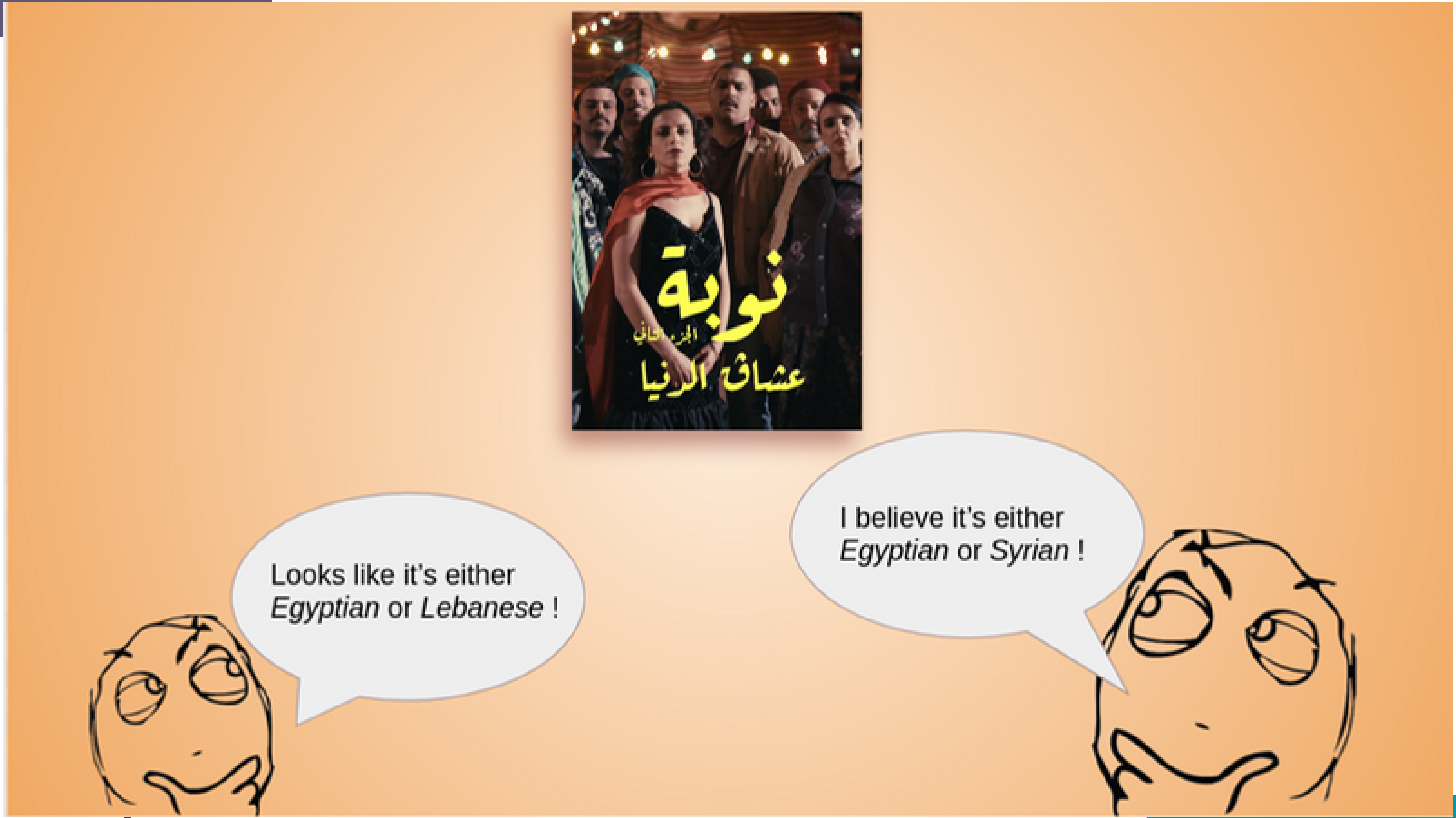
Chap I: Before Dark Times

What reconstructing information meant during my PhD

Quiz Time


I am once again asking
for you TO PARTICIPATE IN QUIZ
imgflip.com


نوبة
الجزء الثاني
عشاق الدنيا
THE ARCHITECT'S DIARY
NETFLIX


نوبة
الجزء الثاني
عشاق الدنيا
Looks like it's either *Egyptian* or *Lebanese* !
I believe it's either *Egyptian* or *Syrian* !


نوبة
الجزء الثاني
عشاق الدنيا


It's *Tunisian*.


Two out of three believed it is *Egyptian*, it must be so !
Unless..

Framework

- A set of m alternatives X: *{Tunisian, Egyptian, Lebanese, Syrian}*
- A hidden (unknown) ground truth alternative $a^* \in X$: *Tunisian*
- A set of n voters N: 3 Friends
- A profile of n approval ballots $A_i \subseteq X$: *{Egy, Leb}, {Egy, Syr}, {Tun}*

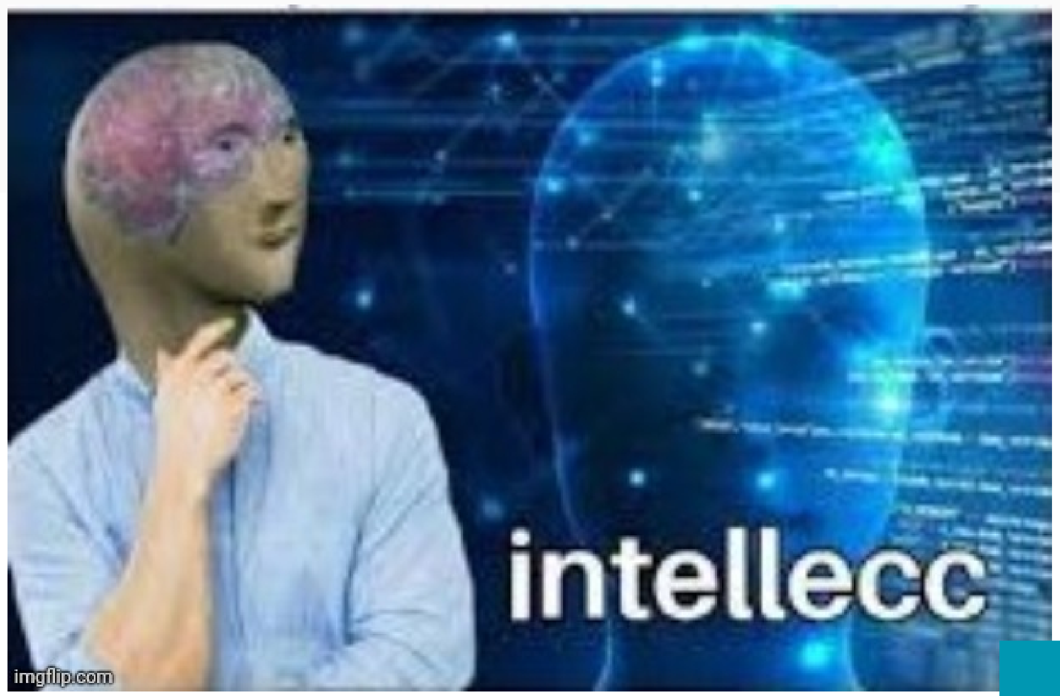
(+) Noise model: probability distribution over the set of possible approval ballots.

$\Longrightarrow$: Estimate the ground truth given the approval ballots by Maximum Likelihood Estimation.

Main Idea

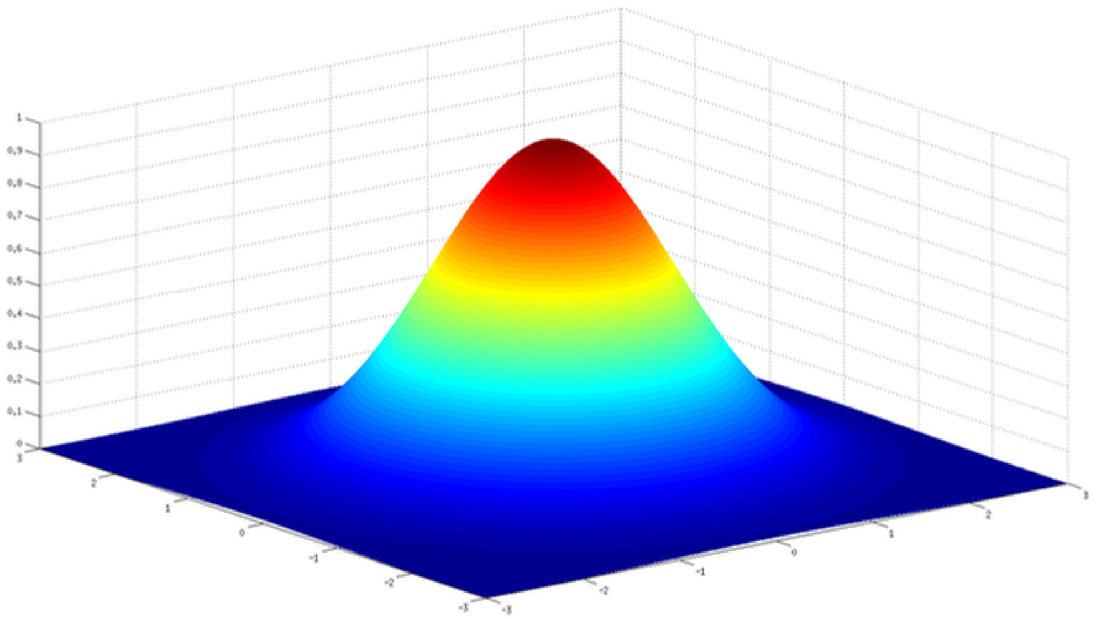
When voters answer truthfully:

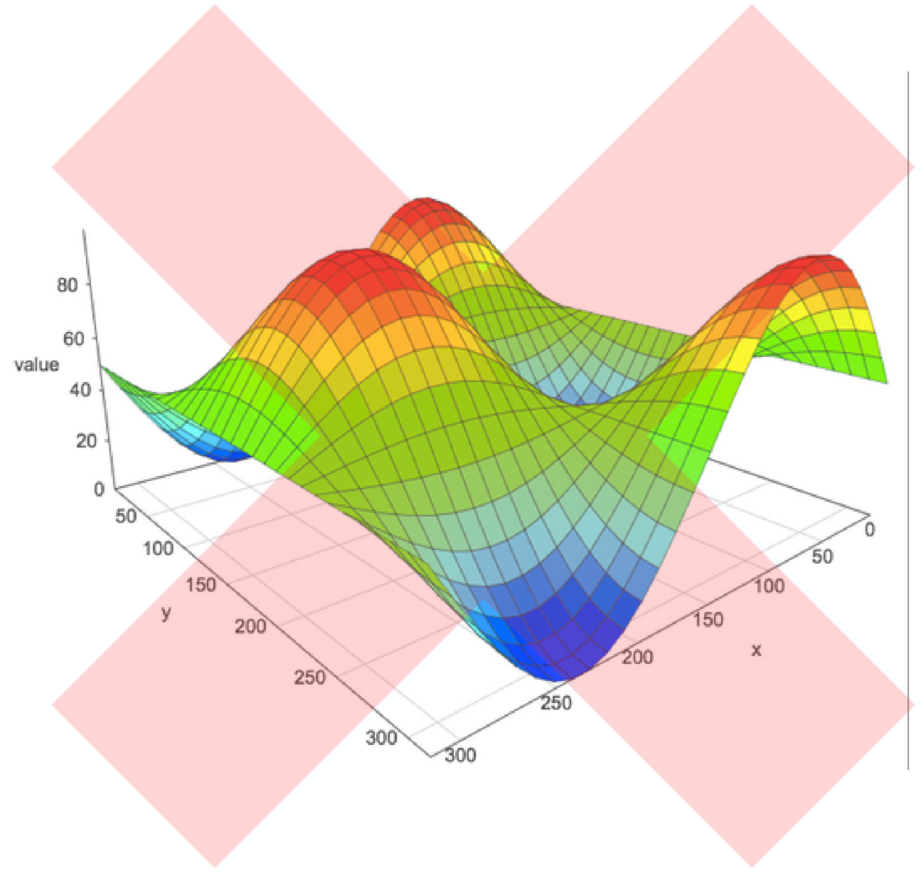
- A voter who knows the correct answer would select a single alternative.
- A voter who selects all the alternatives has no idea of the correct answer.



Noise Models: Mallows

$$P_{\phi_i,d}(A_i|a^* = a) = \frac{1}{Z_i}\phi_i^{d(a^*,A_i)}, \forall a \in \mathcal{X}$$





Neutrality

$$\forall \pi \in \sigma(\mathcal{X}), P_{\phi,d}(A|a^* = a) = P_{\phi,d}(\pi(A)|a^* = \pi(a))$$

$$\Updownarrow$$

$$d(a, A) = \psi_d(|a \cap A|, |A|)$$

Example:

$$d_H(a, A) = |\overline{a} \cap A| + |a \cap \overline{A}| = 1 - 2|a \cap A| + |A|$$

Homogeneous Noise

$$P_{\phi,d}(A_i|a^* = a) = \frac{1}{Z}\phi^{d(a^*,A_i)} = \frac{1}{Z}\phi^{\psi_d(|a^*\cap A_i|,|A_i|)}$$

Theorem

For $n \geq 3$, the maximum likelihood estimation rule ζ_d is a size-decreasing approval rule if and only if:

$$\Delta\psi_d : j \mapsto \psi_d(0,j) - \psi_d(1,j) \text{ is decreasing}$$

Homogeneous Noise

Theorem

For $n \geq 3$, the maximum likelihood estimation rule ζ_d is a size-decreasing approval rule if and only if:

$$\Delta\psi_d : j \mapsto \psi_d(0, j) - \psi_d(1, j) \text{ is } \textcolor{red}{\textit{decreasing}}$$

Examples:

- **Jaccard: 1/card**
- **Dice: 2/(card+1)**

Heterogeneous Noise

$$\psi_d(|a \cap A|, |A|) = f(|a \cap A|) + g(|A|)$$

Theorem

If for every $1 \leq k \leq m - 1$ we have that:

$$g(k+1) - g(k) \geq \frac{1}{2}[f(0) - f(1)]$$

Then:

$$\frac{\partial \mathbb{E}_{\phi,d}[|A_i|]}{\partial \phi} \geq 0$$

Condorcet Noise

$$P_{p_i}(a \in A_i | a = a^*) = P_{p_i}(a \notin A_i | a \neq a^*) = p_i, \forall a \in X$$

Theorem

For $m \geq 2$, we have that:

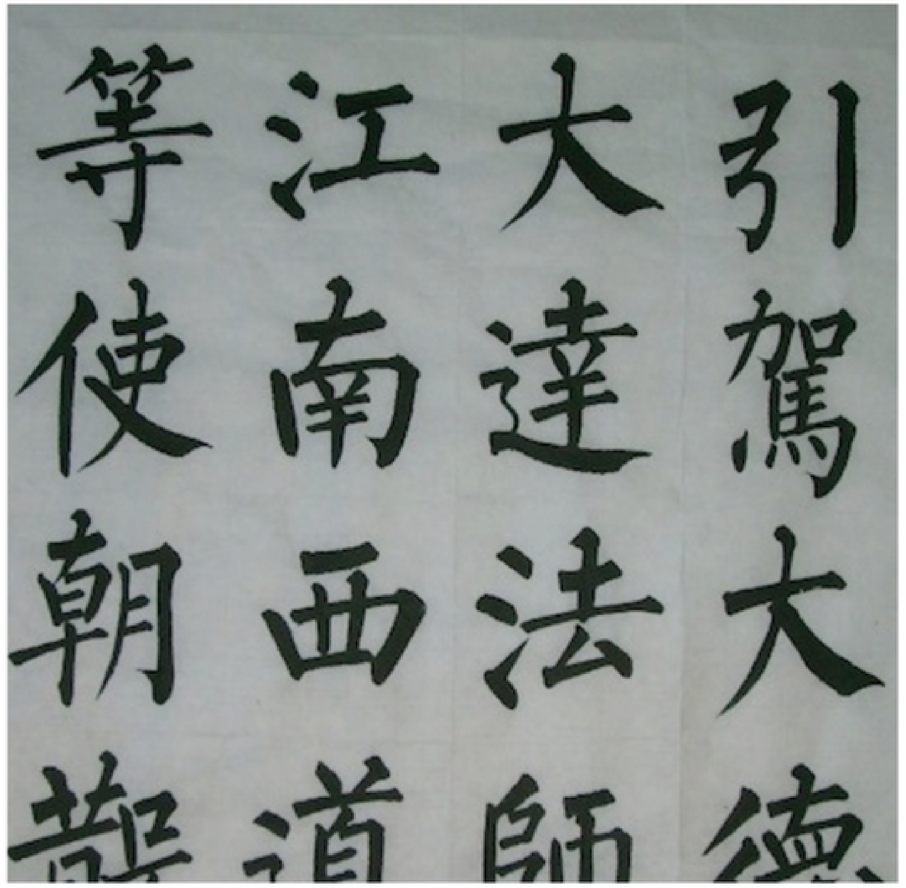
$$\mathbb{E}_p[|A_i|] = (m-1) - (m-2)p$$

Why is this interesting?

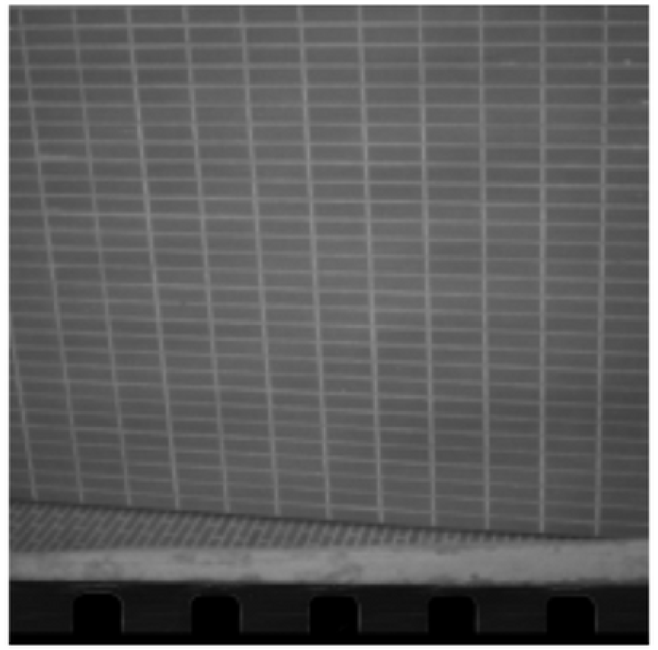
$$\boxed{p_i = \frac{m - 1 - \mathbb{E}_{p_i}[|A_i|]}{m-2}}$$

Experiments

- [x] Leopard
- [ ] Tiger
- [ ] Puma
- [x] Jaguar
- [ ] Lion(ess)
- [x] Cheetah

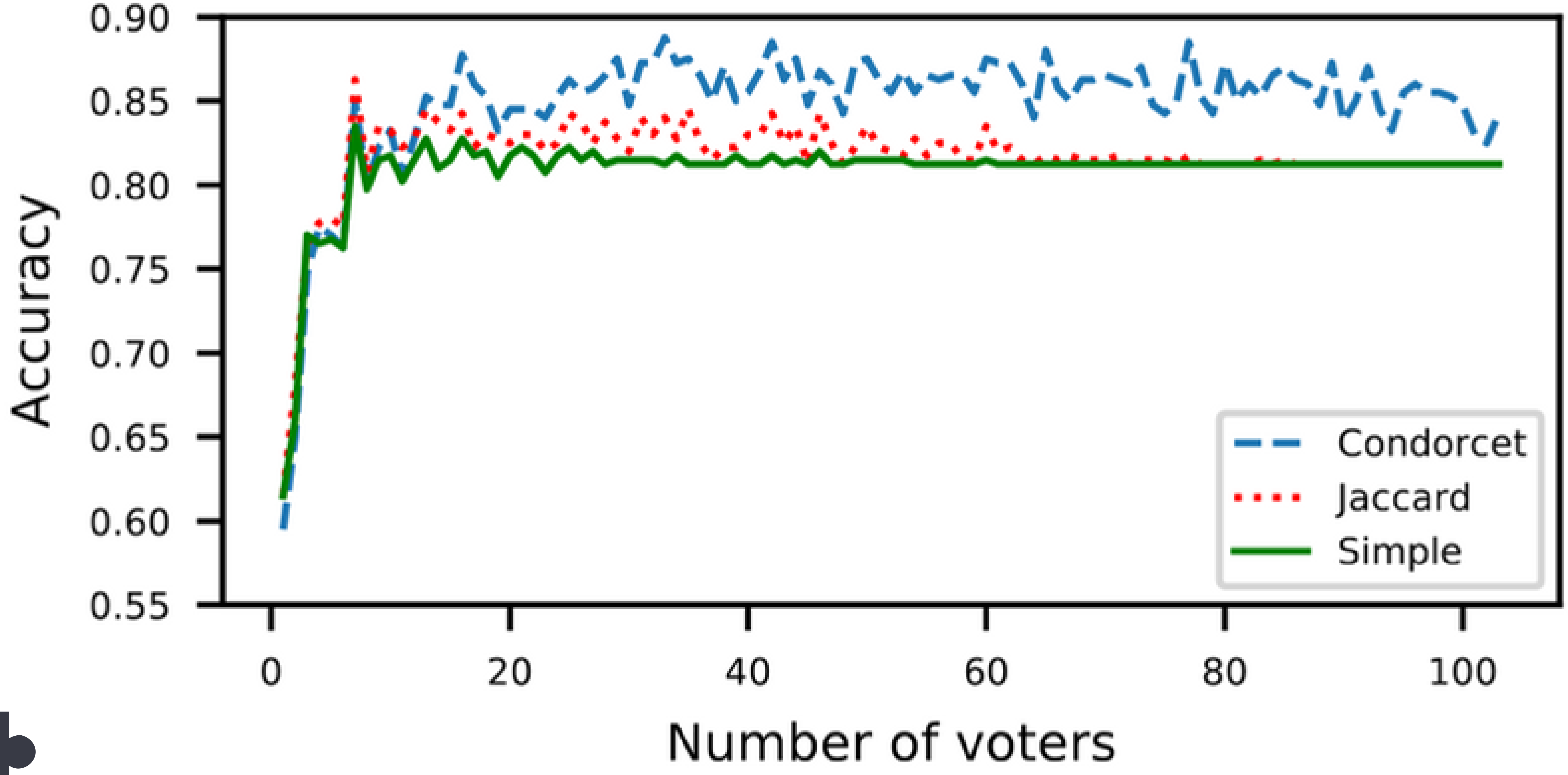


- [ ] Hebrew
- [ ] Russian
- [x] Japanese
- [x] Thai
- [x] Chinese
- [ ] Tamil
- [ ] Latin
- [ ] Hindi



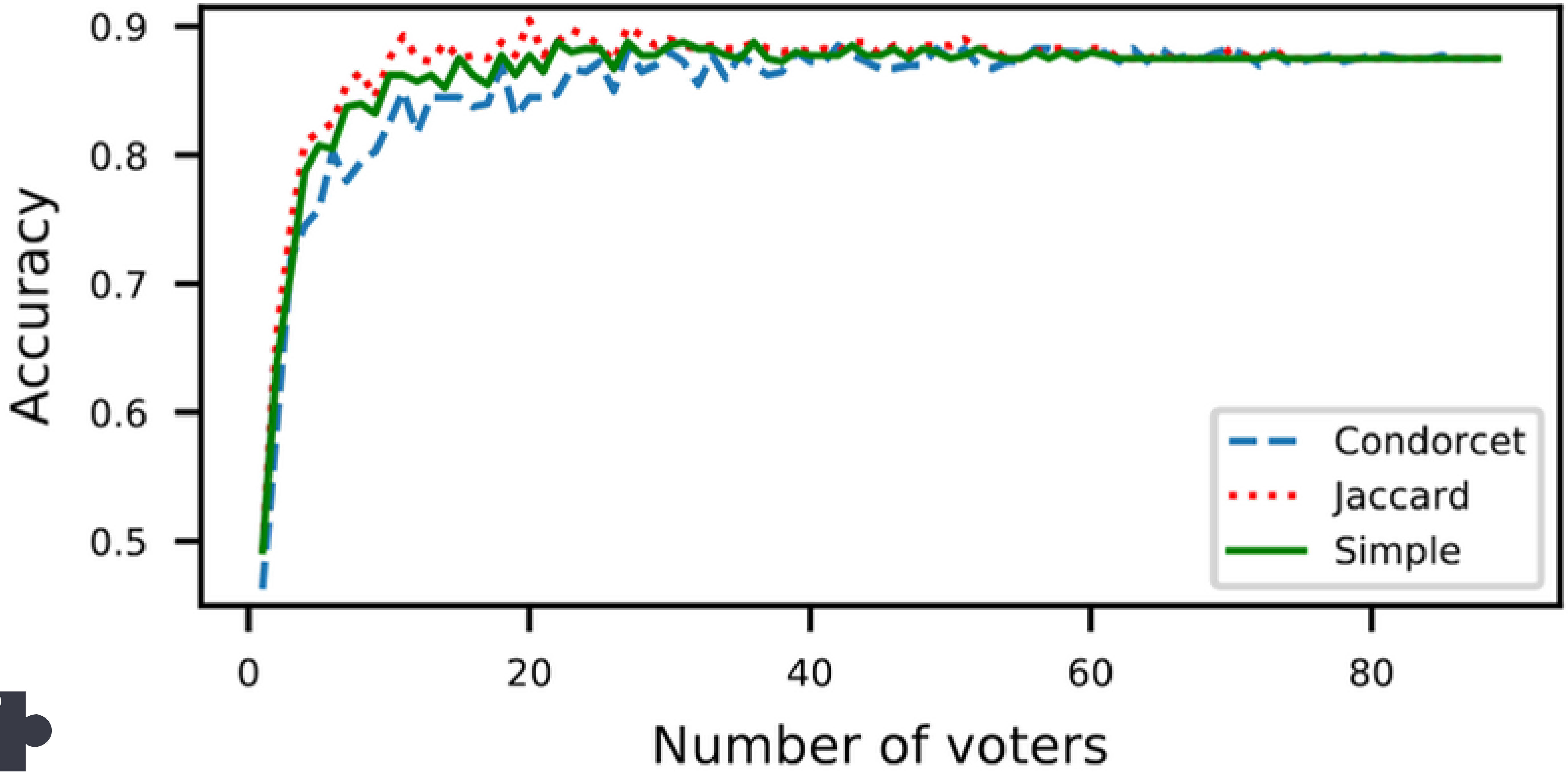
- [x] Gravel
- [ ] Grass
- [x] Brick
- [x] Wood
- [ ] Sand
- [ ] Cloth

Results

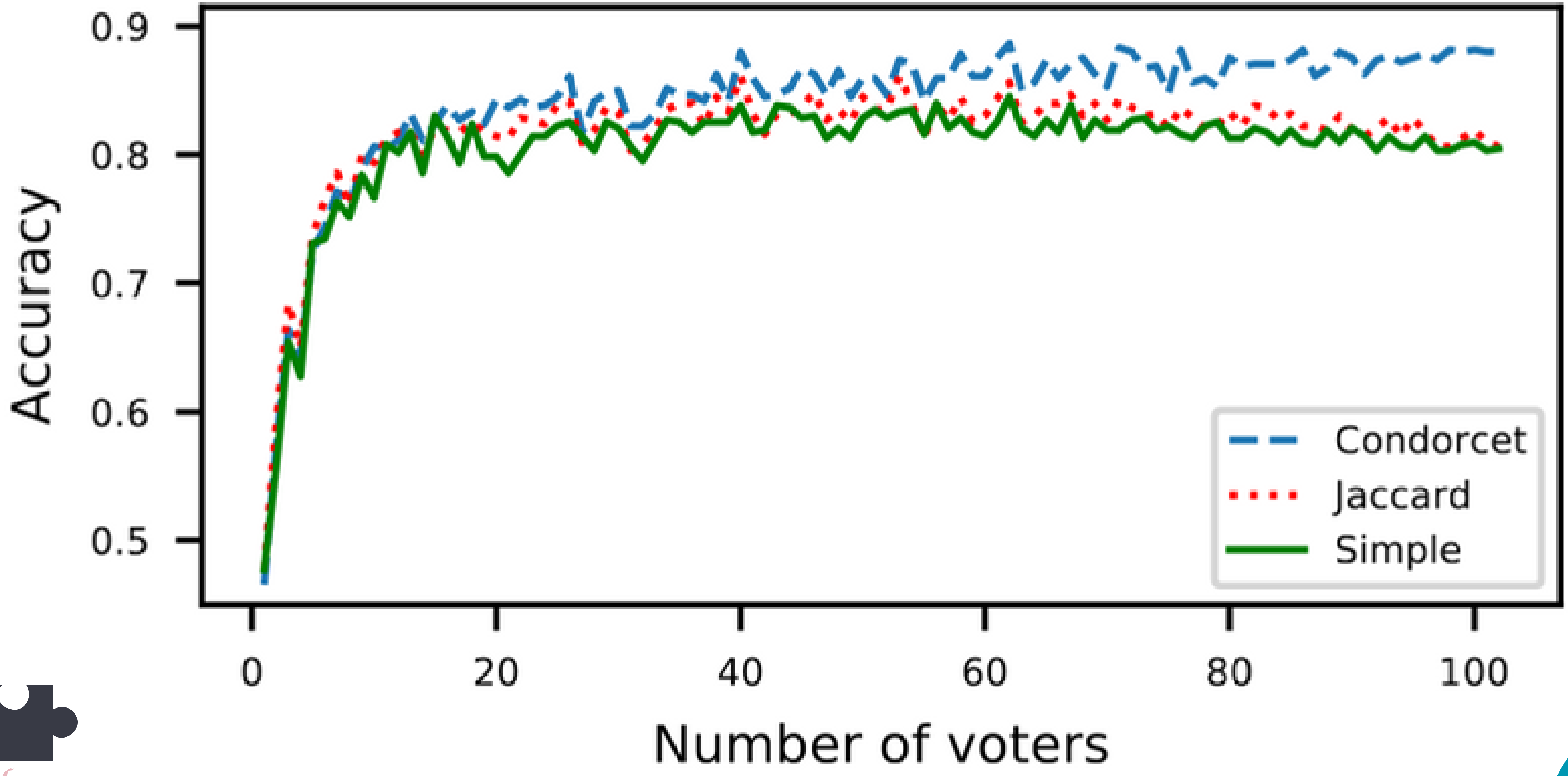


Accuracy
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
20
40
60
80
100
Number of voters
Condorcet
Jaccard
Simple

Results


Accuracy
0.9
0.8
0.7
0.6
0.5
0
20
40
60
80
Number of voters
Condorcet
Jaccard
Simple

Results


Accuracy
0.9
0.8
0.7
0.6
0.5
0
20
40
60
80
100
Number of voters
Condorcet
Jaccard
Simple

つづく

Back to the Quiz..

General Conclusion