

COOPERATIVE DATA ENRICHMENT ALGORITHMS

Tanguy Lefort

IMAG, Univ Montpellier, CNRS

Inria, LIRMM,





- ▶ Alexis Joly
- ▶ Benjamin Charlier
- ▶ Joseph Salmon

- ▶ Pierre Bonnet
- ▶ Antoine Affouard



× *Chitalpa tashkentensis* T.S.Elias & Wisura World flora

Observation



pofpof63
Jun 26, 2023

1: user and date



Most probable name

× *Chitalpa tashkentensis* T.S.Elias & Wisura
Bignoniaceae Dave

2: votes

Submitted name

× *Chitalpa tashkentensis* T.S.Elias & Wisura

Suggested names Vote for the species name

× *Chitalpa tashkentensis* T.S.Elias & Wisura Dave

5



Species name (World flora)



Vote

Badly determined observation? Vote for Undetermined species

⚠ Observation contains pictures of several plants?: Vote for Malformed observation 0



Vesalea grandifolia (Villarreal) Hua Feng Wang & Landrein Flore mondiale Observation

 Pavlos
16 sept. 2023



Nom le plus probable

Vesalea grandifolia (Villarreal) Hua Feng Wang & Landrein
Caprifoliaceae Abélia

Nom soumis

Zabelia triflora (R.Br. ex Wall.) Makino ex Hisauti & H.Hara

Noms suggérés Voter pour le nom d'espèce

Vesalea grandifolia (Villarreal) Hua Feng Wang & L...  3 

Zabelia triflora (R.Br. ex Wall.) Makino ex Hisauti &...  1 

Espèce non identifiée  1 



Espèce (Flore mondiale)



Voter

Observation mal déterminée ? Votez pour Espèce indéterminée



Voter pour un organe



Corrected initial
submission

BUT SOMETIMES USERS CAN'T BE TRUSTED



Espèce non identifiée Flore mondiale

Observation



Nom le plus probable

Espèce non identifiée

Nom soumis

Plantago subulata L.

Noms suggérés Voter pour le nom d'espèce

Plantago subulata L. Plantain à feuilles en alène

👍 5

Espèce non identifiée

👍 2

Polytrichum commune Hedw.

👍 2

Polytrichum commune

👍 1



Espèce (Flore mondiale)



Voter



Voter pour un organe



Voter pour la qualité

Corrected ?

BUT SOMETIMES USERS CAN'T BE TRUSTED



Espèce non identifiée Flore mondiale

Observation



Nom le plus probable

Espèce non identifiée

Nom soumis

Plantago subulata L.

Noms suggérés Voter pour le nom d'espèce

Plantago subulata L. Plantain à feuilles en alène	👍 5	👤
Espèce non identifiée	👍 2	👤
Polytrichum commune Hedw.	👍 2	👤
Polytrichum commune	👍 1	👤

Contributeurs



Majority is wrong

Fermer



Voter pour un organe



Voter pour la qualité



General.

- ▶ The good: Fast, easy, cheap data collection



General.

- ▶ The good: Fast, easy, cheap data collection
- ▶ The bad: Noisy labels with different level skills



General.

- ▶ The good: Fast, easy, cheap data collection
- ▶ The bad: Noisy labels with different level skills
- ▶ The ugly: Very few theory, ad-hoc methods to handle noise from users



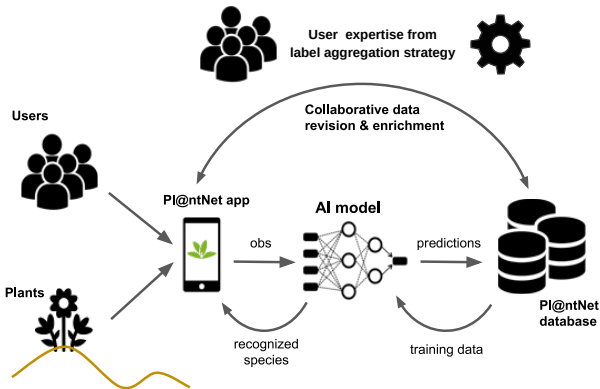
General.

- ▶ The good: Fast, easy, cheap data collection
- ▶ The bad: Noisy labels with different level skills
- ▶ The ugly: Very few theory, ad-hoc methods to handle noise from users

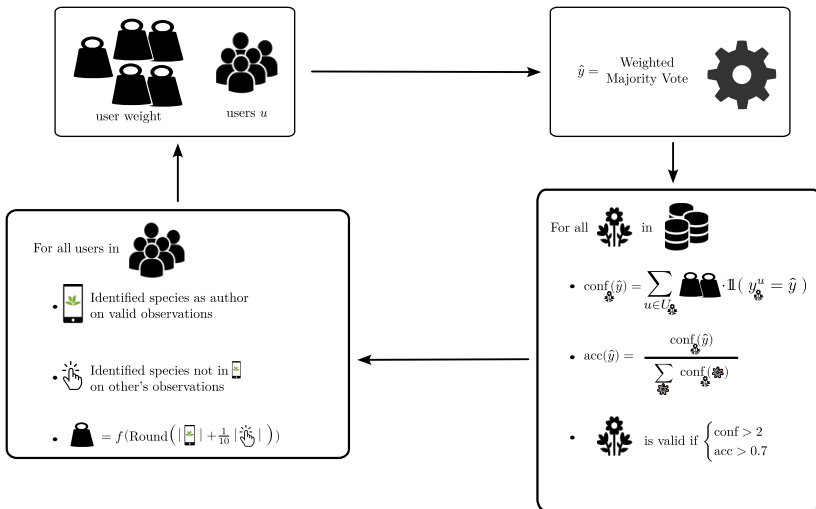
Pl@ntNet.

- ▶ 20+ million observations from around the world
- ▶ 6+ million users
- ▶ 22+ million votes
- ▶ 49 720 species

Key concept of PL@ntNet: Collaborative AI



Weighting users vote by their estimated number of identified species

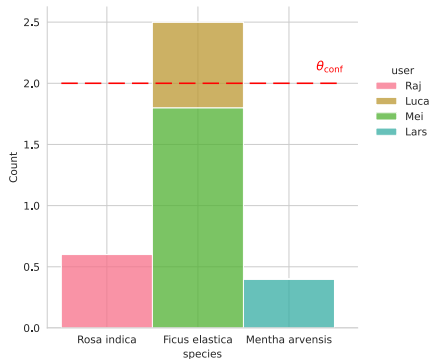
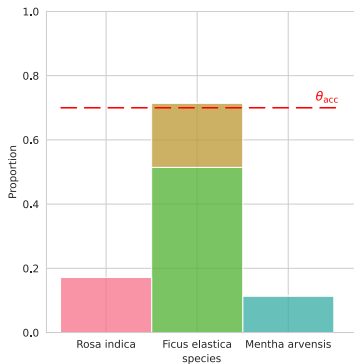


ACTIVE DATASET

ANY OBSERVATION LABELING IS ACTIVE



Initial setting

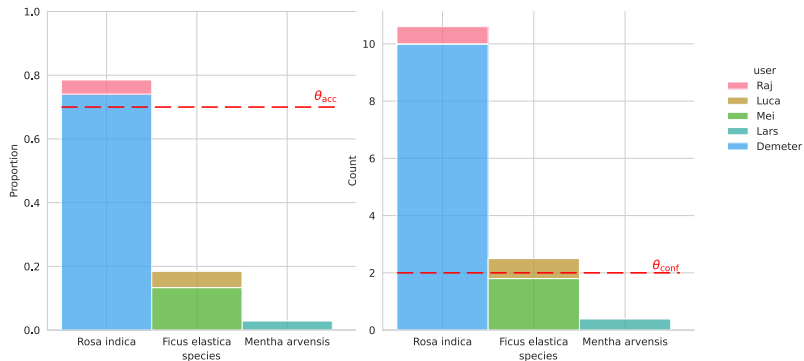


ACTIVE DATASET

ANY OBSERVATION LABELING IS ACTIVE



Label switch

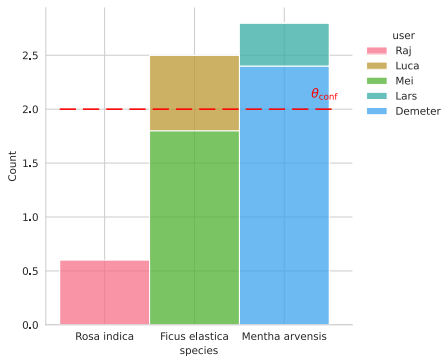
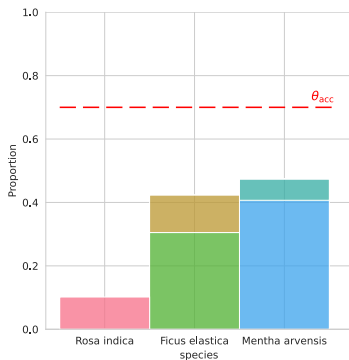


ACTIVE DATASET

ANY OBSERVATION LABELING IS ACTIVE

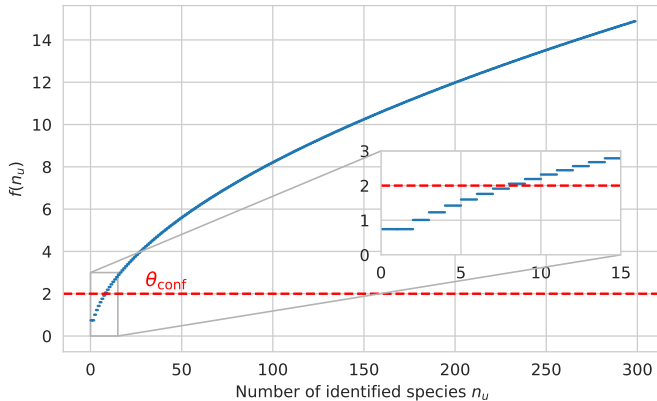


Invalidating label



$$f(n_u) = n_u^\alpha - n_u^\beta + \gamma \text{ with } \begin{cases} \alpha = 0.5 \\ \beta = 0.2 \\ \gamma = \log(1.7) \simeq 0.74 \end{cases}$$

Weight function determination





► **Majority Vote** (MV)



- ▶ **Majority Vote (MV)**
- ▶ **Worker agreement with aggregate (WAWA)** (Appen 2021)
 - ▶ Majority vote
 - ▶ Weight user by how much they agree with the majority
 - ▶ Weighted majority vote



- ▶ **Majority Vote (MV)**
- ▶ **Worker agreement with aggregate (WAWA)** (Appen 2021)
 - ▶ Majority vote
 - ▶ Weight user by how much they agree with the majority
 - ▶ Weighted majority vote
- ▶ **iNaturalist**
 - ▶ Need 2 votes
 - ▶ 2/3 of agreements



- ▶ South Western European flora obs since 2017
- ▶ 823 000 users answered more than 11000 species
- ▶ 6 700 000 observations
- ▶ 9 000 000 votes casted
- ▶ **Imbalance:** 80% of observations are represented by 10% of total votes



- ▶ South Western European flora obs since 2017
- ▶ 823 000 users answered more than 11000 species
- ▶ 6 700 000 observations
- ▶ 9 000 000 votes casted
- ▶ **Imbalance:** 80% of observations are represented by 10% of total votes

No ground truth available to evaluate the strategies

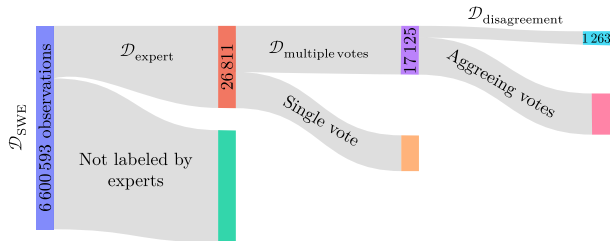
EXTRACTING A SUBSET OF A PL@NTNET

CREATION OF TEST SETS



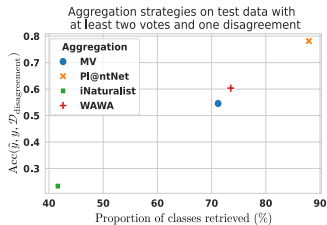
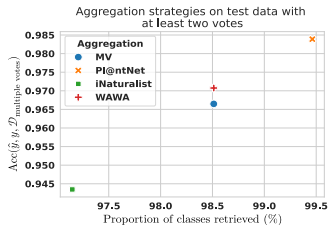
- Extraction of 98 experts (TelaBotanic + prior knowledge – thanks to Pierre Bonnet)

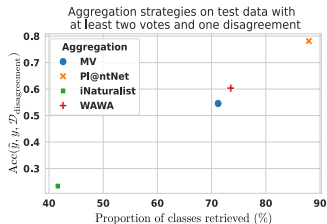
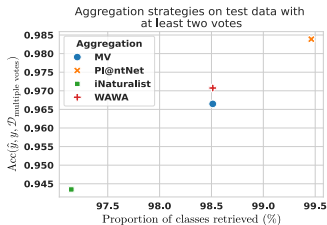
PL@ntnet South-Western Europe flora dataset



PERFORMANCE

ACCURACY AND VOLUME OF CLASSES KEPT



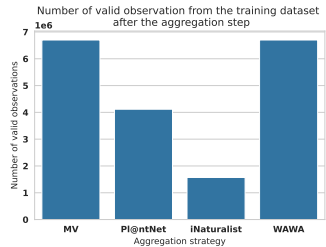
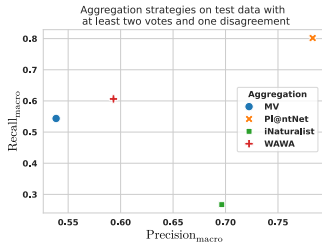


In short

- ▶ PI@ntNet aggregation performs better overall
- ▶ iNaturalist is highly impacted by their reject threshold
- ▶ In ambiguous settings (right), strategies weighting users are better

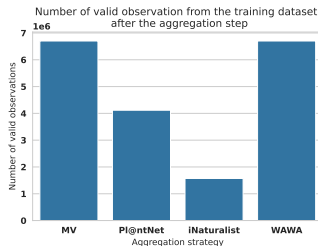
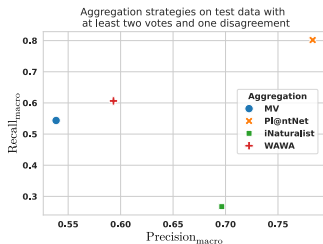
PERFORMANCE

PRECISION, RECALL AND VALIDITY



PERFORMANCE

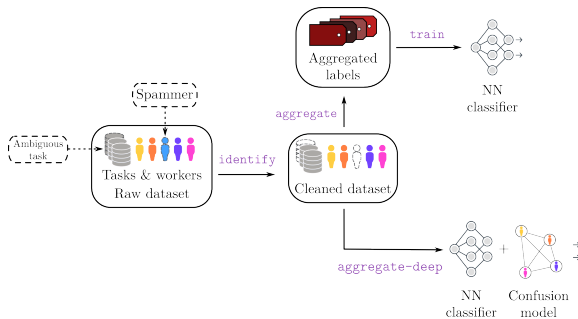
PRECISION, RECALL AND VALIDITY



In short

- ▶ Pl@ntNet aggregation performs better overall
- ▶ iNaturalist has good precision but bad recall
- ▶ We indeed remove some data but less than iNaturalist

Peerannot: Python library to handle crowdsourced data



Questions?
Cookies?



Why?

- ▶ More data
- ▶ Could correct non expert users
- ▶ Could invalidate bad quality data



Why?

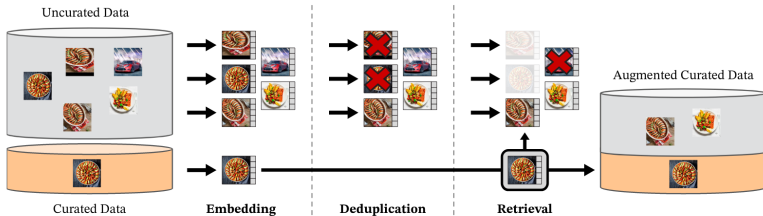
- ▶ More data
- ▶ Could correct non expert users
- ▶ Could invalidate bad quality data

Dangers

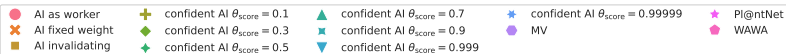
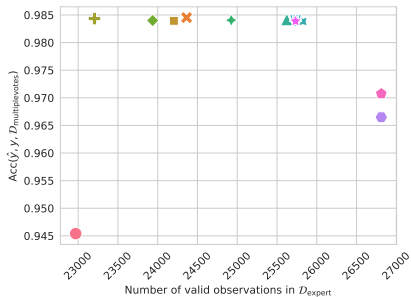
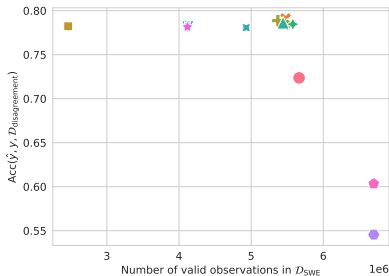
- ▶ Redundancy: users are already guided by AI prediction
- ▶ Model collapse from training on its generated data
- ▶ If the network acts as a control agent, who controls the network?

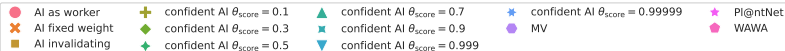
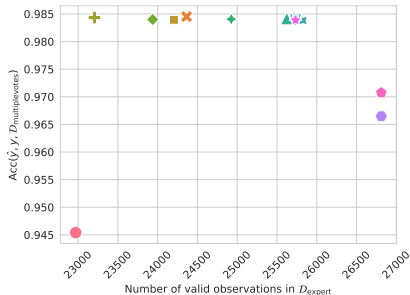
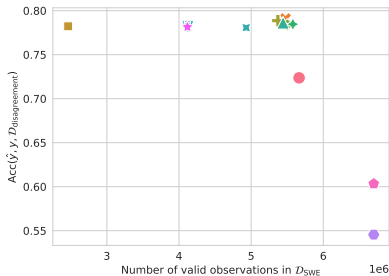
- ▶ AI **as worker**: naive integration
- ▶ AI **fixed weight**: weight= 1.7 to invalidate two new users, but $< \theta_{\text{conf}}$
- ▶ AI **invalidating**: fixed weight but can only invalidate observations
- ▶ AI **confident**: fixed weight on data with $\mathbb{P}(\text{predicted species}) > \theta_{\text{score}}$

DinoV2 (Oqab et. al 2024) trained monthly (transformers based)



PERFORMANCE COMPARISON





In short

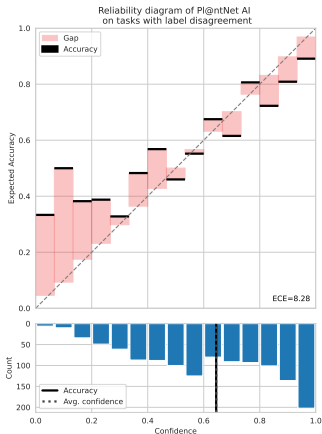
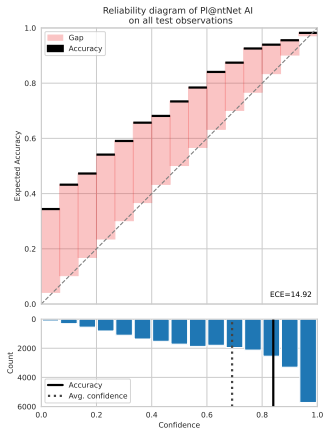
- ▶ AI should not be considered as any other user
- ▶ More stable results: **confident AI** with $\theta_{\text{score}} = 0.7$

NOTE ON CALIBRATION

OVER OR UNDERCONFIDENCE?



If we use probability outputs: can they be considered as probabilities?



Aggregation strategy

- ▶ Pl@ntNet aggregation fits the large scale framework
- ▶ With a system to invalidate data and clean the training set

Aggregation strategy

- ▶ Pl@ntNet aggregation fits the large scale framework
- ▶ With a system to invalidate data and clean the training set

AI vote

- ▶ Confident AI seems the best performing
- ▶ We should calibrate the network before deployment

Thank you!

