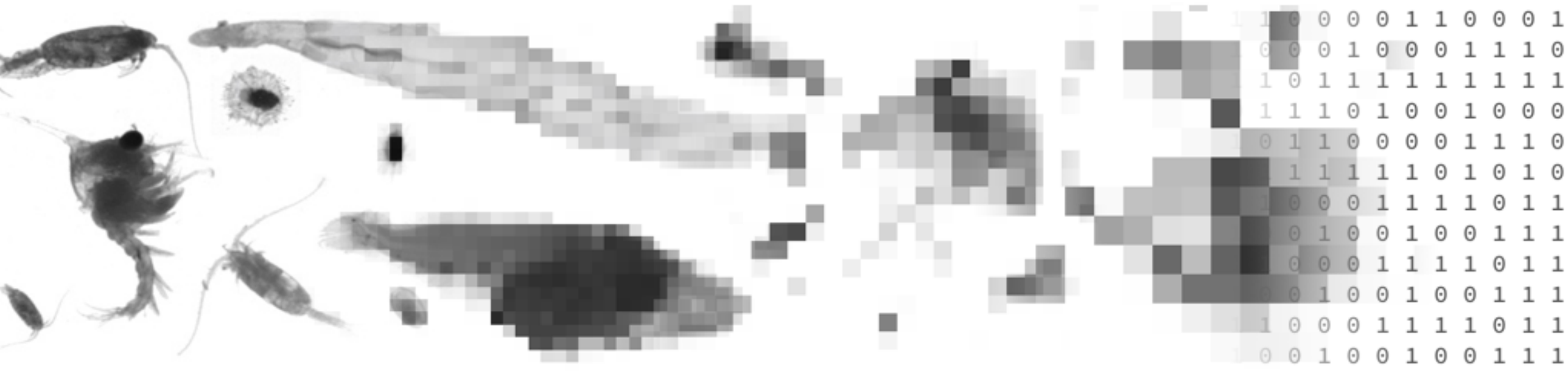
WKMLEARN, 2018-04-18
Jean-Olivier Irisson, Martin Schröder, Marc Picheral

EcoTaxa

A human-computer interface to classify images along a taxonomy with the help of machine-learning



Why EcoTaxa at the Laboratoire d'Océanographie de Villefranche (LOV)?

Why EcoTaxa at the Laboratoire d'Océanographie de Villefranche (LOV)?

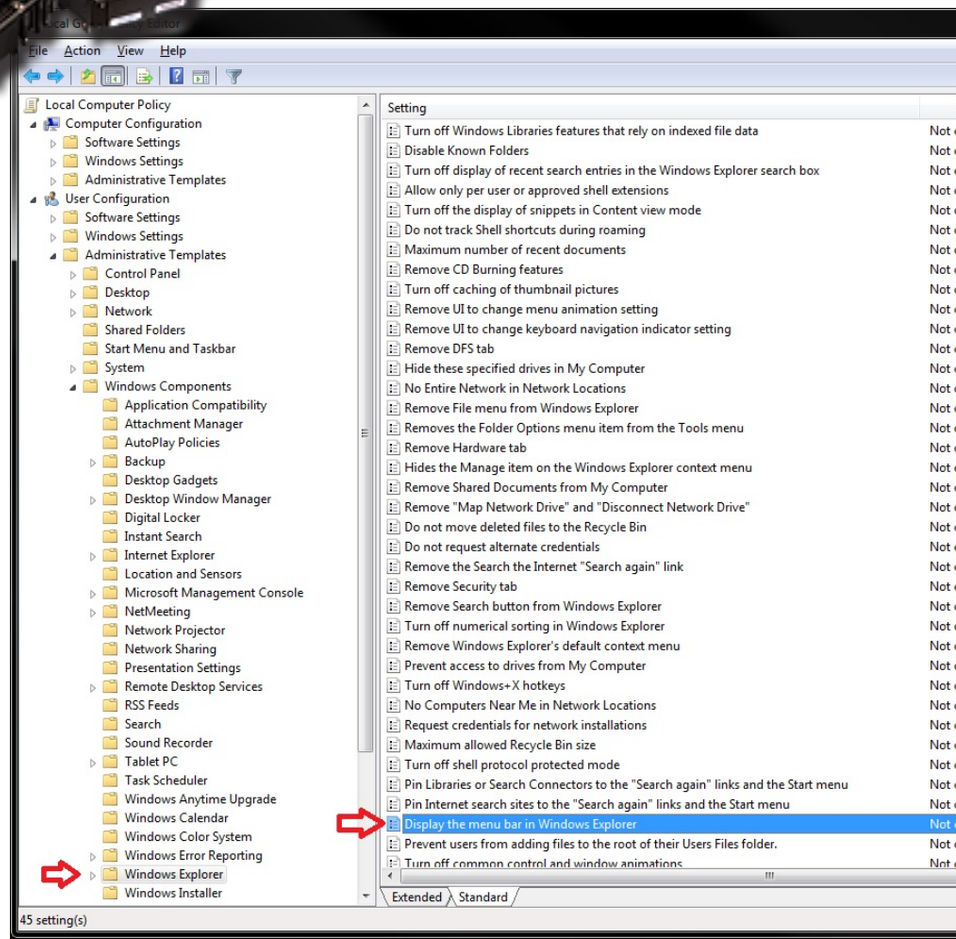
Why EcoTaxa at the Laboratoire d'Océanographie de Villefranche (LOV)?

Why EcoTaxa at the Laboratoire d'Océanographie de Villefranche (LOV)?

Data management

Streamlining of the user interaction (and in particular the use of machine learning models)

Data accessibility and collaboration



Demo

http://ecotaxa.obs-vlfr.fr

http://ecotaxa.obs-vlfr.fr/prj/859

http://ecotaxa.obs-vlfr.fr/prj/857

http://ecotaxa.obs-vlfr.fr/prj/858

Data model

PROJECT
title
status
mapping
settings

SAMPLE
operator
cruise
depth
volume
...

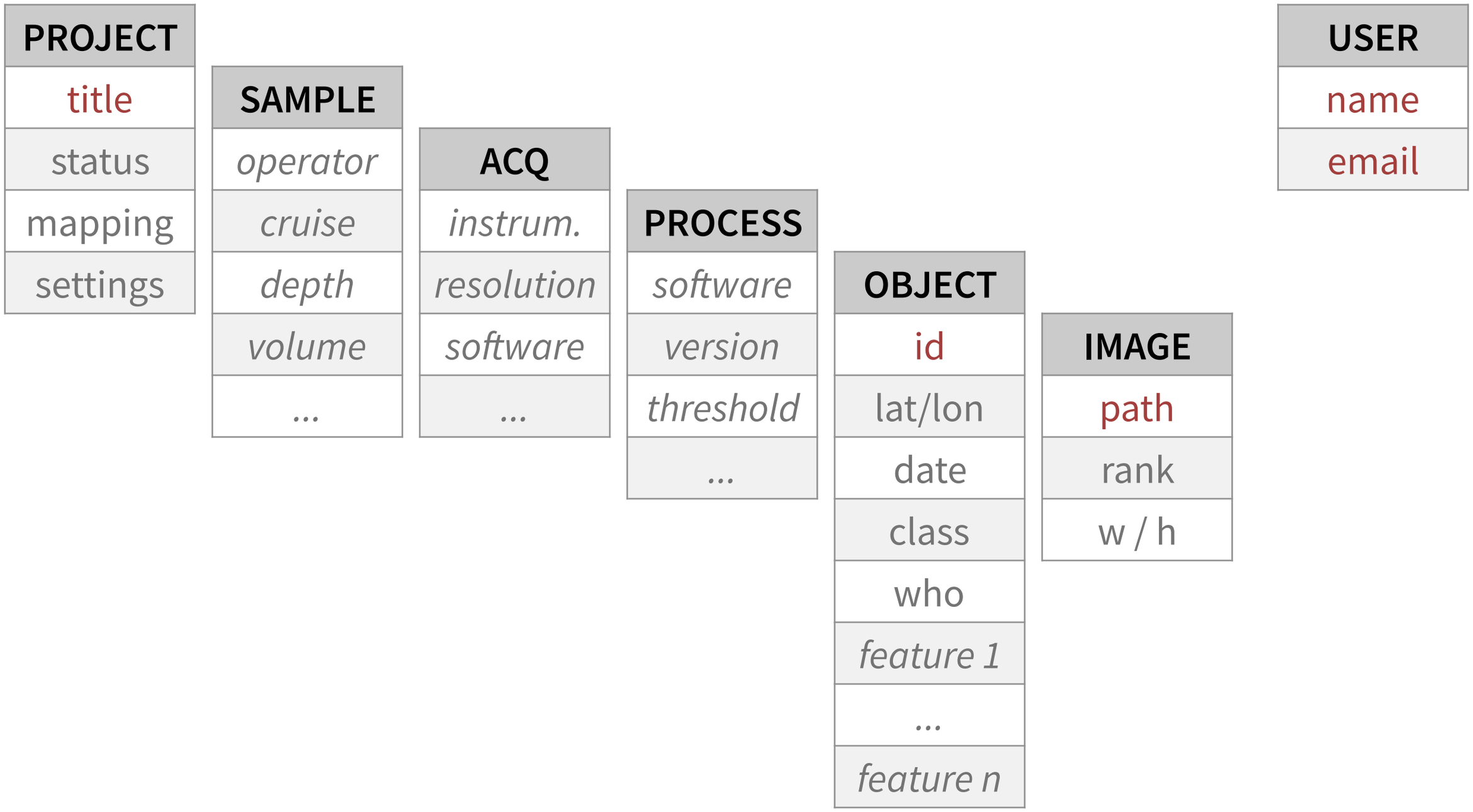
ACQ
instrum.
resolution
software
...

PROCESS
software
version
threshold
...

OBJECT
id
lat/lon
date
class
who
feature 1
...
feature n

IMAGE
path
rank
w / h

Data model


PROJECT
title
status
mapping
settings
SAMPLE
operator
cruise
depth
volume
...
ACQ
instrum.
resolution
software
...
PROCESS
software
version
threshold
...
OBJECT
id
lat/lon
date
class
who
feature 1
...
feature n
USER
name
email
IMAGE
path
rank
w / h

Data model

PROJECT
title
status
mapping
settings

SAMPLE
operator
cruise
depth
volume
...

ACQ
instrum.
resolution
software
...

PROCESS
software
version
threshold
...

OBJECT
id
lat/lon
date
class
who
feature 1
...
feature n

IMAGE
path
rank
w / h

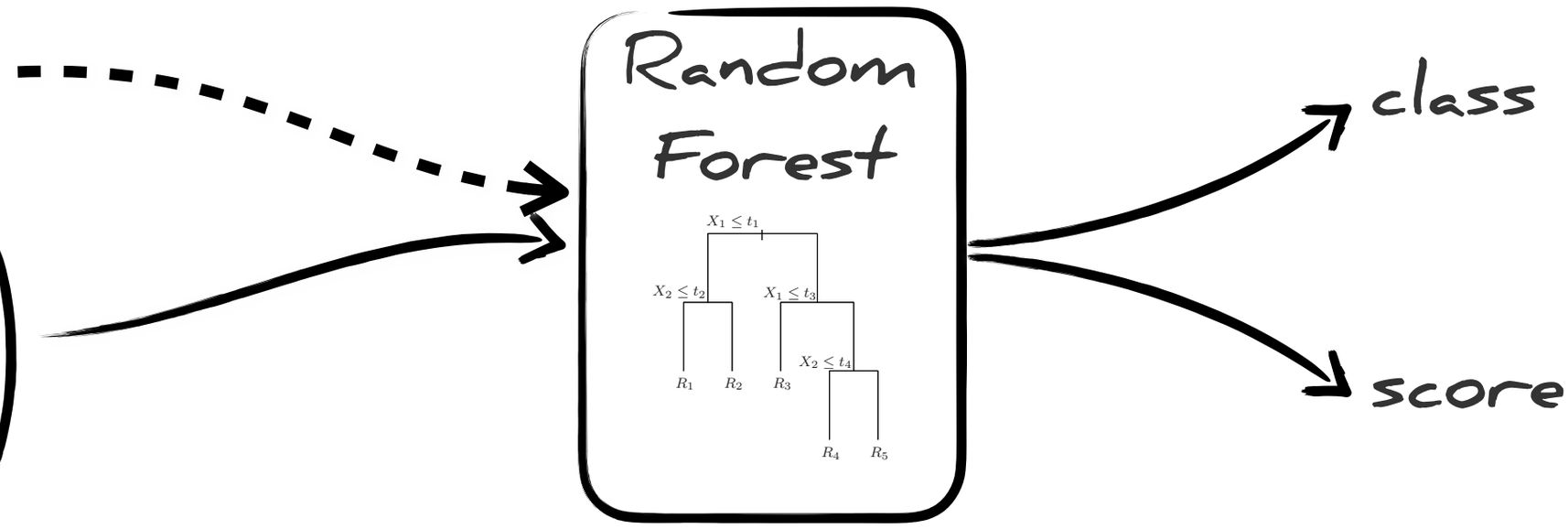
USER
name
email

TAXO
id
name
parent_id

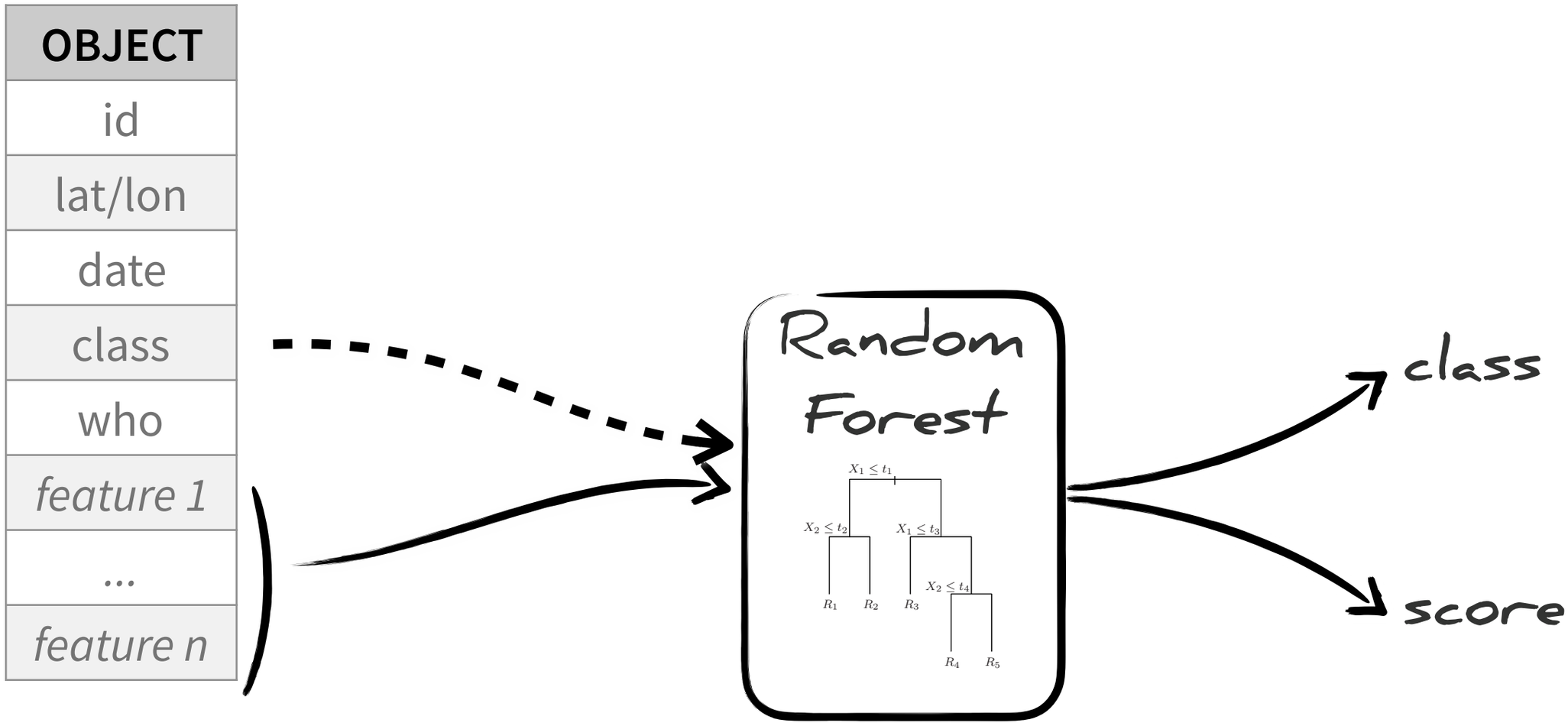
https://unieuk.org

Machine learning workflow

OBJECT
id
lat/lon
date
class
who
feature 1
...
feature n

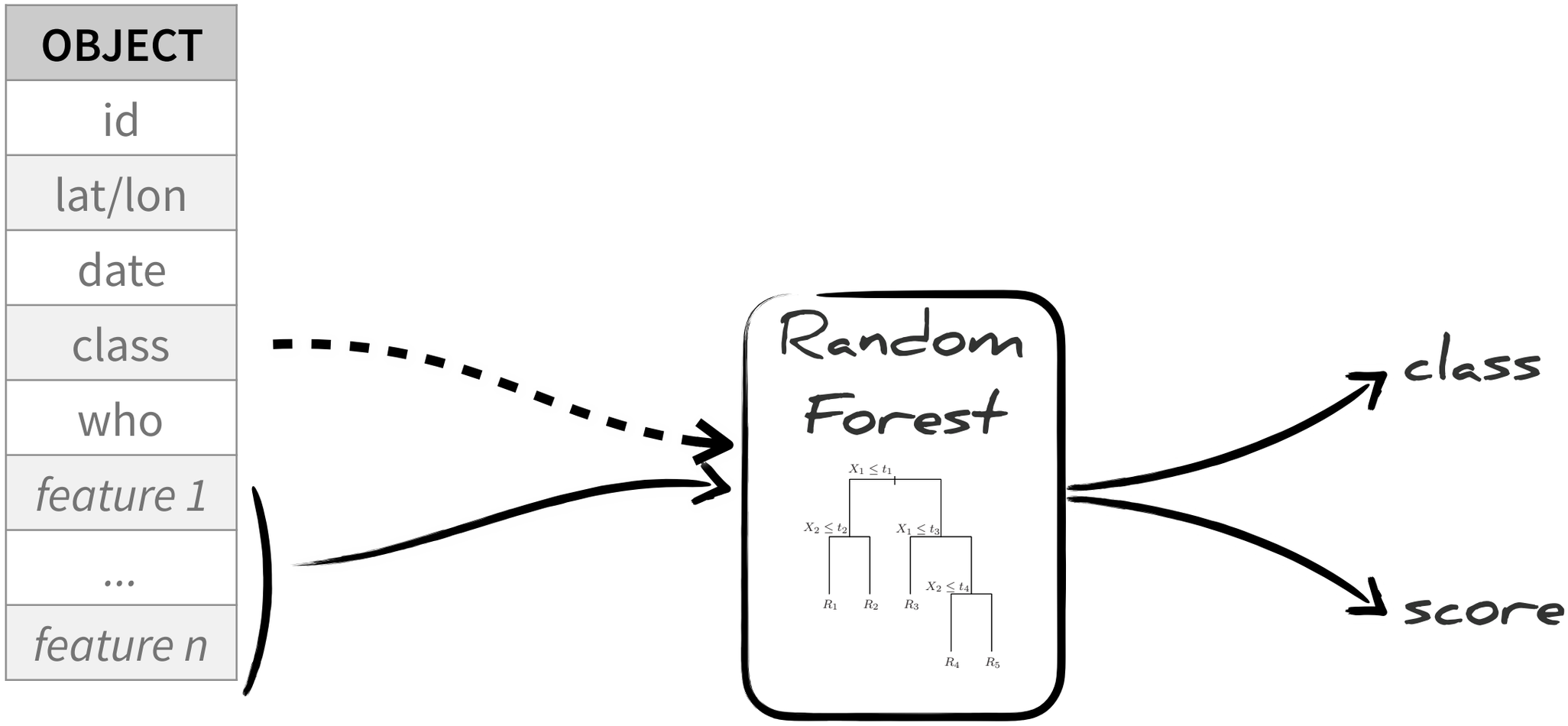


Machine learning workflow



Dataset	Grouping	Random Forest (ZP)	Random Forest (SCN)	Random Forest (ZP+SCN)	SparseConvNet
flowcam	group1	82.65%	82.98%	86.64%	69.22%
flowcam	group2	82.56%	83.05%	86.54%	68.15%
uvp5ccelter	group1	85.89%	79.96%	86.70%	52.36%
uvp5ccelter	group2	85.91%	79.74%	86.63%	51.68%
zoocam	group1	87.49%	90.30%	92.16%	88.64%
zoocam	group2	92.74%	93.46%	95.31%	89.85%
zooscan	group1	71.14%	78.90%	79.96%	61.09%
zooscan	group2	70.84%	78.22%	79.62%	60.58%

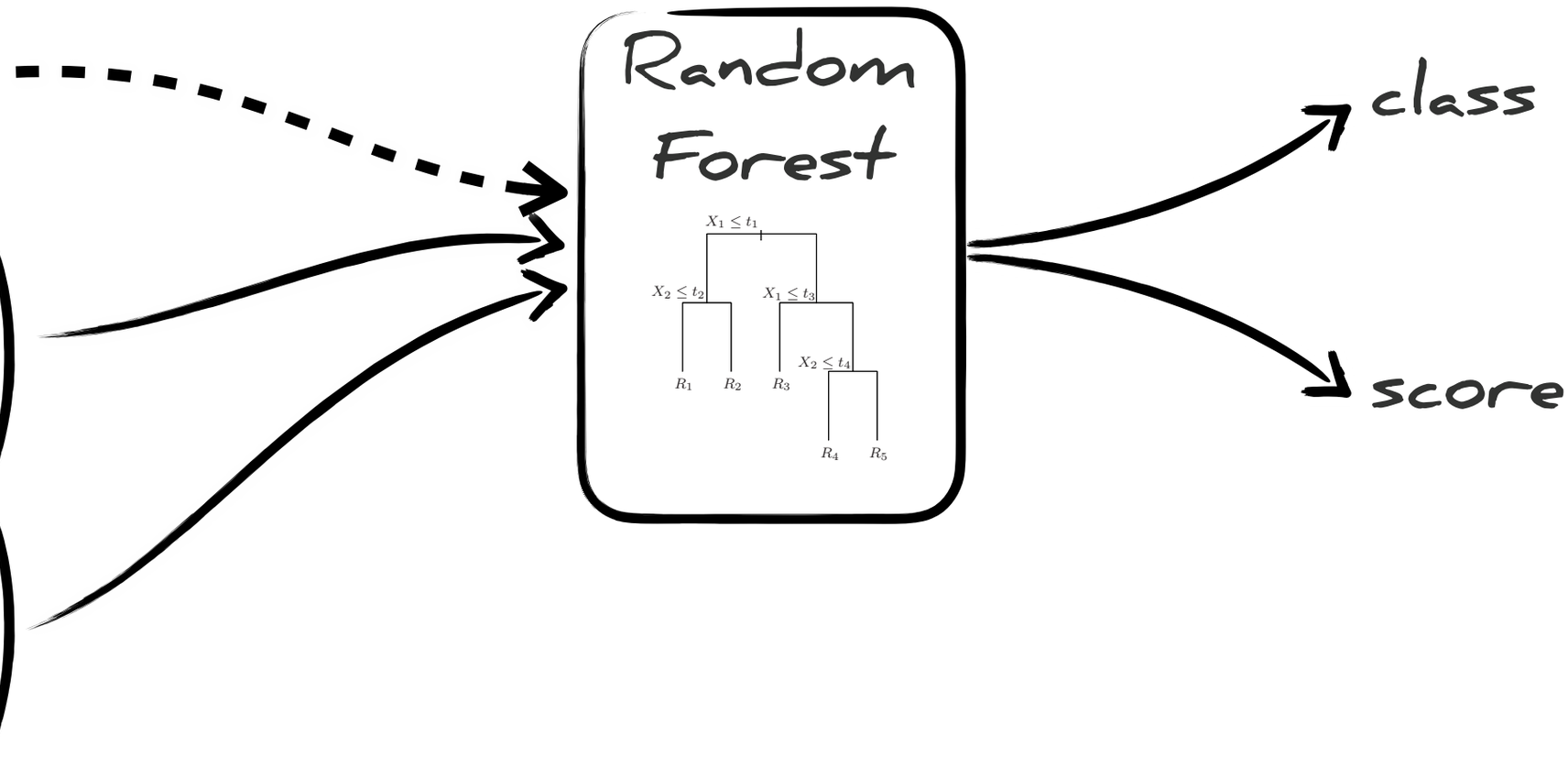
Machine learning workflow



Dataset	Grouping	Random Forest (ZP)	Random Forest (SCN)	Random Forest (ZP+SCN)	SparseConvNet
flowcam	group1	82.65%	82.98%	86.64%	69.22%
flowcam	group2	82.56%	83.05%	86.54%	68.15%
uvp5ccelter	group1	85.89%	79.96%	86.70%	52.36%
uvp5ccelter	group2	85.91%	79.74%	86.63%	51.68%
zoocam	group1	87.49%	90.30%	92.16%	88.64%
zoocam	group2	92.74%	93.46%	95.31%	89.85%
zooscan	group1	71.14%	78.90%	79.96%	61.09%
zooscan	group2	70.84%	78.22%	79.62%	60.58%

Machine learning workflow

OBJECT
id
lat/lon
date
class
who
feature 1
...
feature n
cnn 1
...
cnn 50



Why SparseConvNet?

Sparsity

faster, more efficient

varying input size

Factional Max Pooling

https://www.kaggle.com/c/datasciencebowl

Why SparseConvNet?

Sparsity

faster, more efficient

varying input size

Factional Max Pooling

https://www.kaggle.com/c/datasciencebowl

kaggle Search kaggle Competitions Datasets Kernels Discussion Learn ••• Sign In

National Data Science Bowl

Predict ocean health, one plankton at a time

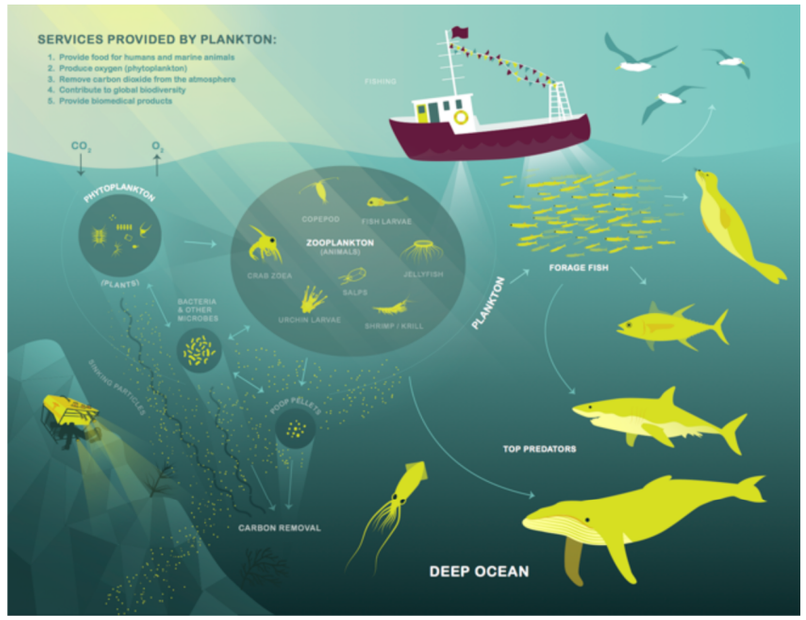
$175,000 · 1,049 teams · 3 years ago

Overview Data Discussion Leaderboard Rules

Overview

Description
Evaluation
Prizes
About The Ndsb
Timeline
Tutorial

Plankton are critically important to our ecosystem, accounting for more than half the primary productivity on earth and nearly half the total carbon fixed in the global carbon cycle. They form the foundation of aquatic food webs including those of large, important fisheries. Loss of plankton populations could result in ecological upheaval as well as negative societal impacts, particularly in indigenous cultures and the developing world. Plankton's global significance makes their population levels an ideal measure of the health of the world's oceans and ecosystems.



Traditional methods for measuring and monitoring plankton populations are time consuming and cannot scale to the granularity or scope necessary for large-scale studies. Improved approaches are needed. One such approach is through the use of an underwater imagery sensor. This towed, underwater camera system captures microscopic, high-resolution images over large study areas. The images can then be analyzed to assess species populations and distributions.

Manual analysis of the imagery is infeasible – it would take a year or more to manually analyze the imagery volume captured in a single day. Automated image classification using machine learning tools is an alternative to the manual approach. Analytics will allow analysis at speeds and scales previously thought impossible. The automated system will have broad applications for assessment of ocean and ecosystem health.

The National Data Science Bowl challenges you to build an algorithm to automate the image identification process. Scientists at the Hatfield Marine Science Center and beyond will use the algorithms you create to study marine food webs, fisheries, ocean conservation, and more. This is your chance to contribute to the health of the world's oceans, one plankton at a time.

Acknowledgements

Booz | Allen | Hamilton & kaggle

The National Data Science Bowl is presented by
with data provided by the Hatfield Marine Science Center at Oregon State University.

Why SparseConvNet?

Sparsity

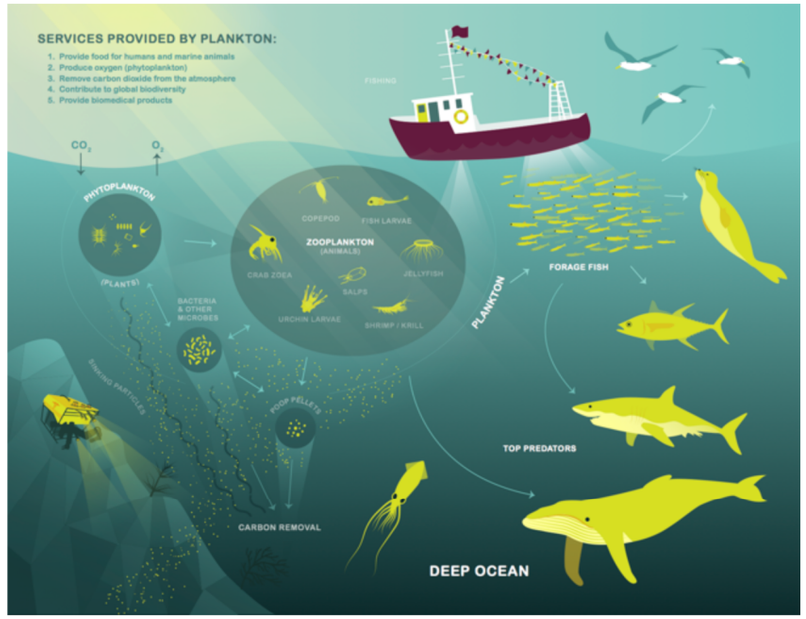
faster, more efficient

varying input size

Factional Max Pooling

https://www.kaggle.com/c/datasciencebowl

on earth and nearly half the total carbon fixed in the global carbon cycle. They form the foundation of aquatic food webs including those of large, important fisheries. Loss of plankton populations could result in ecological upheaval as well as negative societal impacts, particularly in indigenous cultures and the developing world. Plankton's global significance makes their population levels an ideal measure of the health of the world's oceans and ecosystems.



Traditional methods for measuring and monitoring plankton populations are time consuming and cannot scale to the granularity or scope necessary for large-scale studies. Improved approaches are needed. One such approach is through the use of an underwater imagery sensor. This towed, underwater camera system captures microscopic, high-resolution images over large study areas. The images can then be analyzed to assess species populations and distributions.

Manual analysis of the imagery is infeasible – it would take a year or more to manually analyze the imagery volume captured in a single day. Automated image classification using machine learning tools is an alternative to the manual approach. Analytics will allow analysis at speeds and scales previously thought impossible. The automated system will have broad applications for assessment of ocean and ecosystem health.

The National Data Science Bowl challenges you to build an algorithm to automate the image identification process. Scientists at the Hatfield Marine Science Center and beyond will use the algorithms you create to study marine food webs, fisheries, ocean conservation, and more. This is your chance to contribute to the health of the world's oceans, one plankton at a time.

Acknowledgements

Booz | Allen | Hamilton & kaggle

The National Data Science Bowl is presented by
with data provided by the Hatfield Marine Science Center at Oregon State University.

Leaderboard >

1 ≈ Deep Sea ≈
2 Happy Lantern Festival
3 Poisson Process
4 Junonia
5 ⚓Deepsea Challenger⚓
6 AuroraXie
7 Maxim Milakov
8 Ilya Kostrikov

155 discussion topics >

Error on submission
2 replies · 2 months ago

scikit-learn Random Forest memory problem
9 replies · 2 months ago

Code to print out predictions (probs) for Caffe
13 replies · a year ago

Why use logarithm to evaluate results
0 replies · a year ago

CNN + Caffe
12 replies · a year ago

Network structure in SparseConvNet

```
0:Convolution 2^2x1->4
1:Learn 4->32 VeryLeakyReLU
2:Pseudorandom overlapping Fractional Max Pooling 1.41421 2
... ×12
36:Convolution 2^2x384->1536
37:Learn 1536->416 VeryLeakyReLU
38:TerminalPooling 32 1024
39:Learn 416->448 VeryLeakyReLU
40:Learn 448->93 Softmax Classification

Spatially sparse CNN with layer sizes: 1-(TP)-32-(C2)-33-
(POFMP)-47-(C2)-48-(POFMP)-68-(C2)-69-(POFMP)-98-(C2)-99-
(POFMP)-140-(C2)-141-(POFMP)-199-(C2)-200-(POFMP)-283-
(C2)-284-(POFMP)-402-(C2)-403-(POFMP)-570-(C2)-571-
(POFMP)-808-(C2)-809-(POFMP)-1144-(C2)-1145-(POFMP)-1619-
(C2)-1620-(POFMP)-2291-(C2)-2292
Input-field dimensions = 2292x2292
```

Usage statistics (as of today)

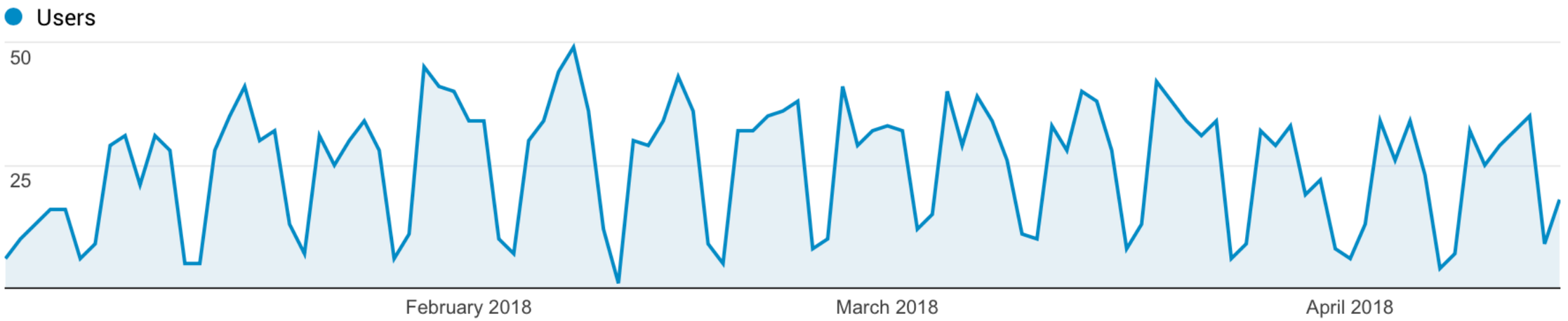
50M objects

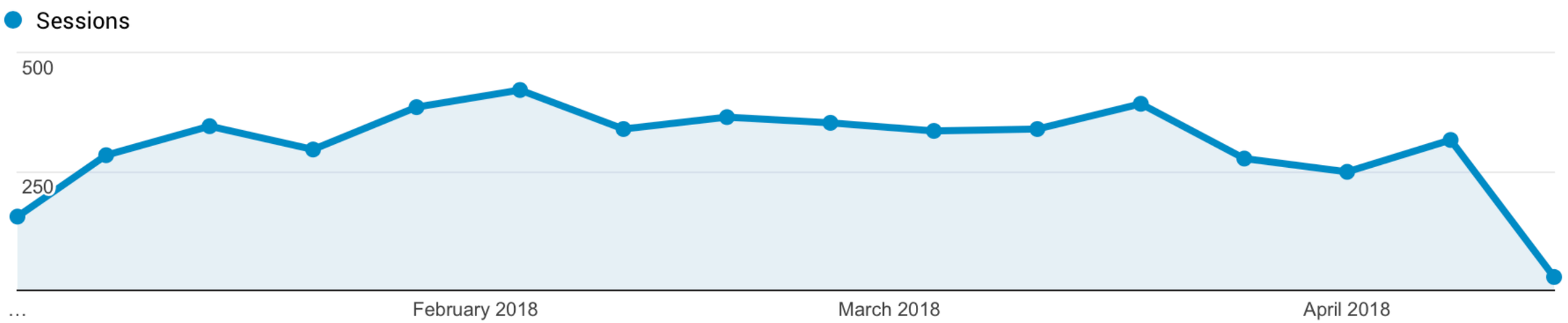
520 projects

350 registered users

92k taxa, 1200 actually used

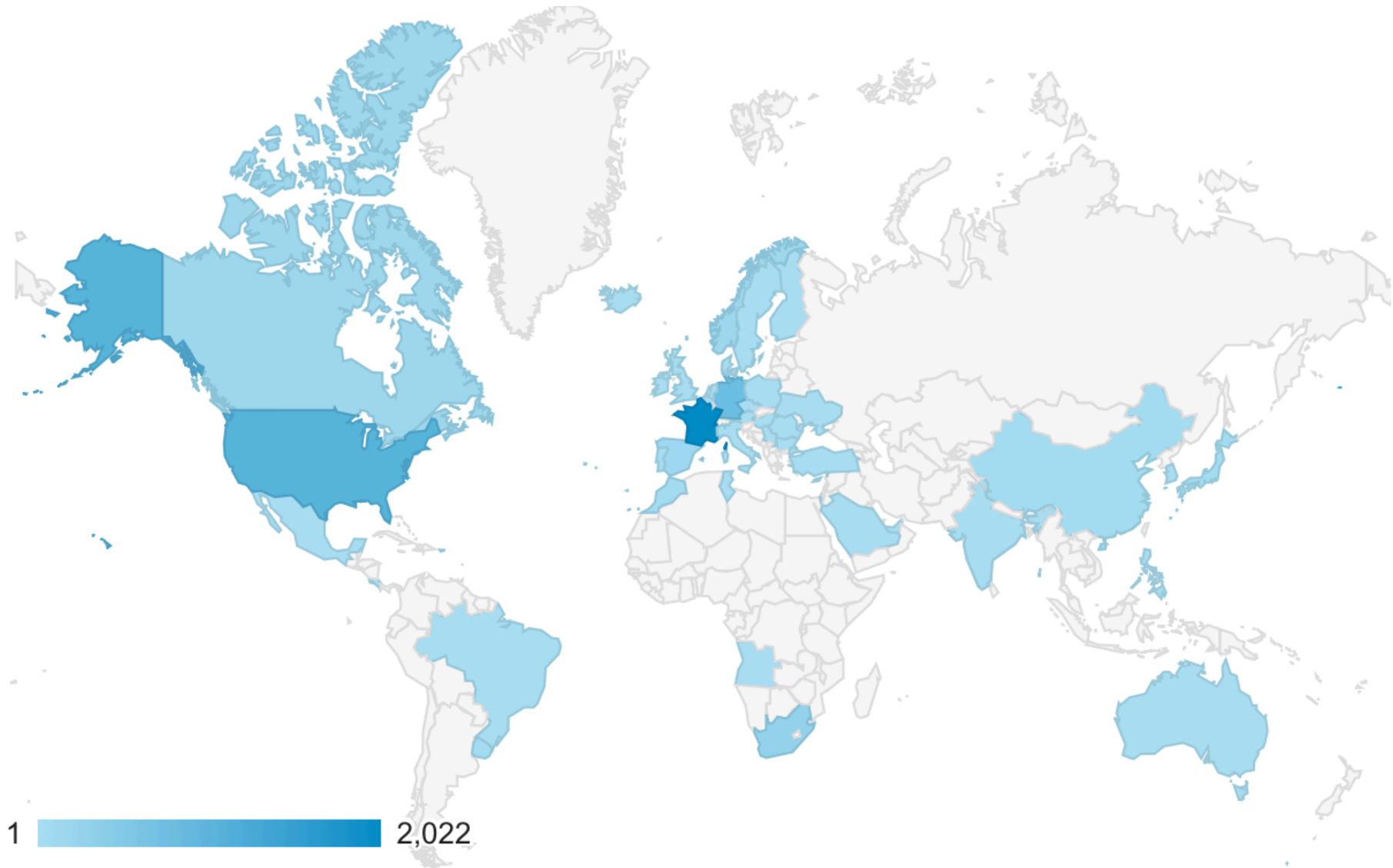
10 to 20k classifications/user/day





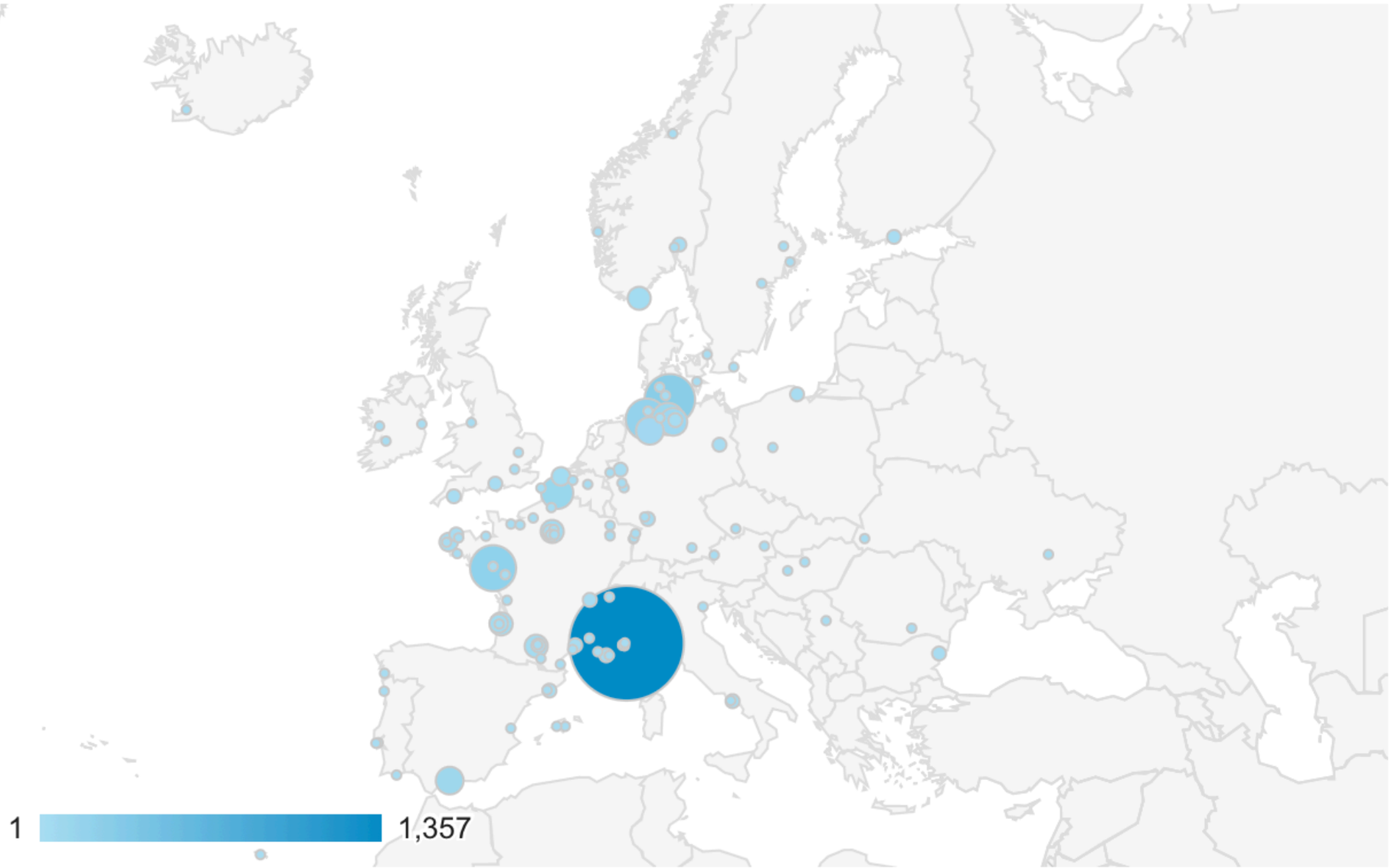
Usage statistics (as of today)

Sessions



Usage statistics (as of today)

Sessions



Future plans for EcoTaxa 1.5

Application

Fix our 99 github **issues** ;-)

Add project-level **stats** (who classified, when, how much, etc.)

Centralised taxonomy and user management

Choose a **taxonomy**!

Centralised image browsing (master vs child instances)

UI overhaul

API

Machine learning

Merge classes before training

More classic **framework** (pytorch?)

Generic trained network (ImageNet)

Use **hierarchy** in taxonomy

Unsupervised **clustering** of large/diverse groups

Continuous training and prediction through active learning

+ classification **embedded** on remote devices (0.1W per 3Mpx image)

How?

Various **collaborations** ongoing

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S)

Facebook Artificial Intelligence Research (FAIR)

Google France

SAP

BG7-A call, proposal coordinated by Ketil!

Future plans for EcoTaxa 1.5

Application

Fix our 99 github **issues** ;-)

Add project-level **stats** (who classified, when, how much, etc.)

Centralised taxonomy and user management

Choose a **taxonomy**!

Centralised image browsing (master vs child instances)

UI overhaul

API

Machine learning

Merge classes before training

More classic **framework** (pytorch?)

Generic trained network (ImageNet)

Use **hierarchy** in taxonomy

Unsupervised **clustering** of large/diverse groups

Continuous training and prediction through active learning

+ classification **embedded** on remote devices (0.1W per 3Mpx image)

Merci