SFEcologie 2022, 2022-11-22
Irisson, Salinas, Colin, COMPLEx team, Picheral

SCIENCES SORBONNE UNIVERSITÉ

EcoTaxa

a tool to support the taxonomic classification of large image datasets through supervised machine learning

Why EcoTaxa?

ZooScan

1.5M objects/y

FlowCam

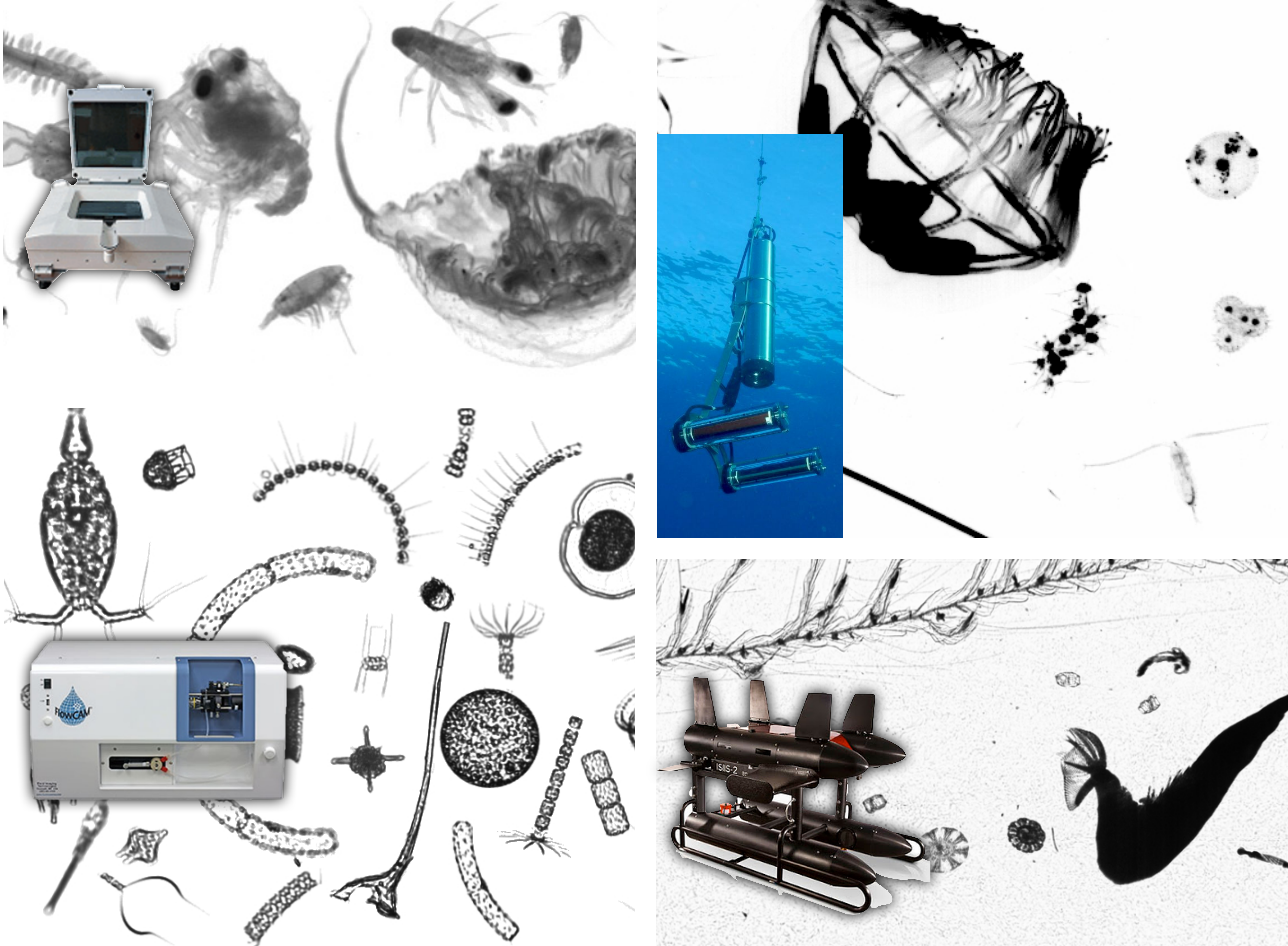
~5M objects/y

UVP

~10M objects/y

ISIIS

100M objects/y



How?

Step 1: Import images (and metadata) into a project

How?

Step 1: Import images (and metadata) into a project

Demo Zooscan JO low taxo res (241 , 1156 , 0 , 0 / 1397) Done :2

Open in a separate window (right click to copy link) Close

Project: **Demo Zooscan JO low taxo res** (managed by : Jean-Olivier Irisson)
To report a mistake, contact Jean-Olivier Irisson

Classification :
fiber<detritus
fiber < detritus < not-living (id=85076)
validated by Jean-Olivier Irisson (irisson@obs-vlfr.fr) on 2022-11-21 20:37:38

Complementary information (edit):
Image list : 1

Set a new classification : Save as Validated Save as dubious Enable Editing Close

Object details | Sample details | Acquisition details | Processing details | Classification change log | Map | Edit complementary informations

longitude	7.31567	**latitude**	43.68500	**Date**	2018-02-05	**Time (daytime)**	12:00:00 (Day)
Depth min	0.0	**Depth max**	75.0	**Classif auto**	Copepoda (0.762)	**Classif auto when**	2021-07-29 14:09:52.442648
Object #	83060411	**Original Object ID**	wp220180205_d2_1_537				
lat_end		**lon_end**		**area**	1918	**mean**	166
stddev	50.73	**mode**	243	**min**	78.00	**max**	248
x	63.21	**y**	61.49	**xm**	63.33	**ym**	61.95
perim.	384	**bx**	4189	**by**	11923	**width**	133
height	99.00	**major**	112	**minor**	21.80	**angle**	140
circ.	0.16	**feret**	156	**intden**	318630	**median**	163

A note on the data architecture

PROJECT
title
settings
CC BY
mapping

SAMPLE
lat/lon
date/time
volume
operator
instrument
resolution
...

SUBSAMP
fraction
px size
frac.method
frac. mesh
...

OBJECT
objID
taxoID
who
when
area
major ax.
minor ax.
feature 4
...
feature n

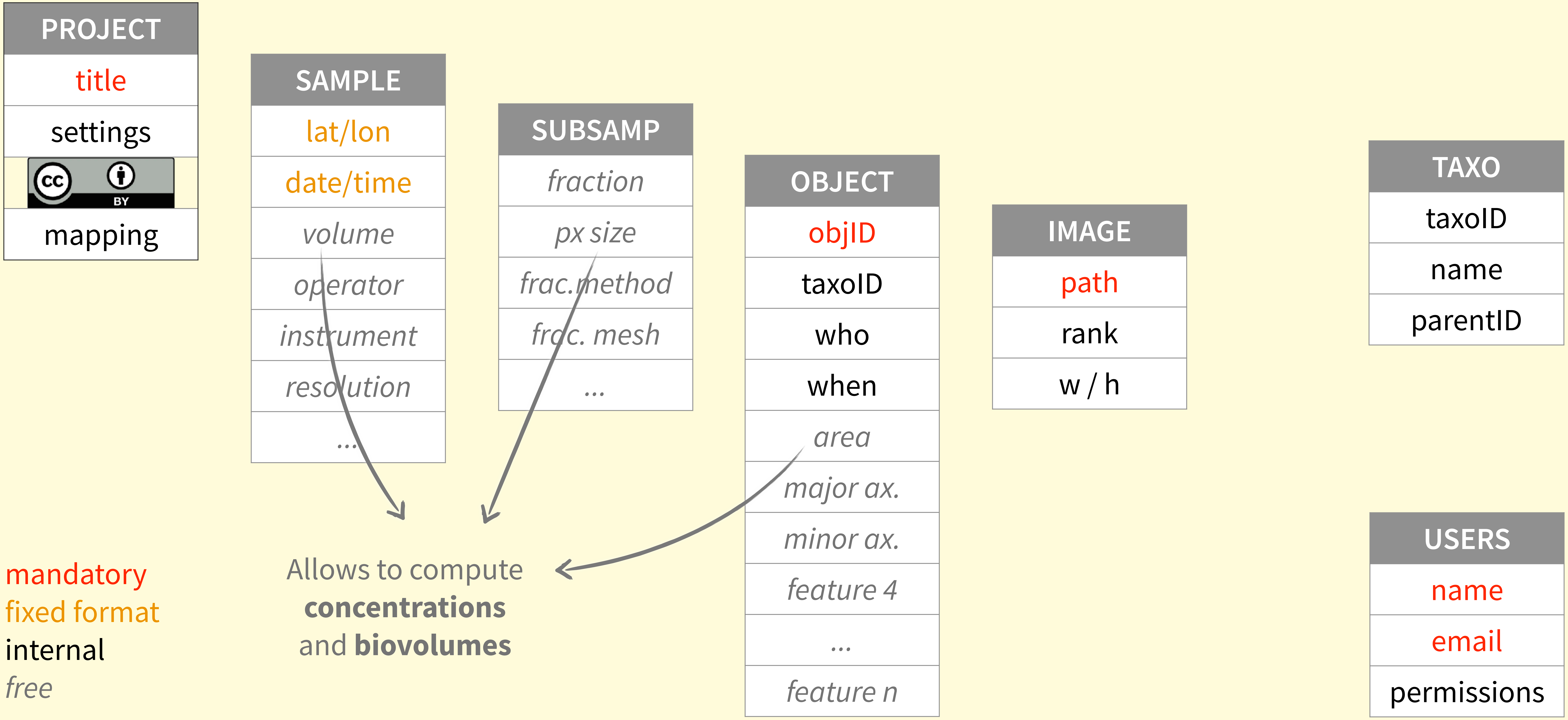
IMAGE
path
rank
w / h

TAXO
taxoID
name
parentID

USERS
name
email
permissions

mandatory
fixed format
internal
free

A note on the data architecture



How?

Step 2: Classify a few images

How?

Step 2: Classify a few images

How?

Step 3: Train a classifier based on classified images and predict the others

How?

Step 3: Train a classifier based on classified images and predict the others

EcoTaxa 2.6 Demo Zooscan JO low taxo res

Jean-Olivier Irisson (log out)

Action

PREDICTION: Choice of Learning Set data source

Next: Choose objects in selected projects

A Learning Set AKA Training Data Set is built using validated data from one or several projects.
Inside these projects, you will be able, in next pages, to specify:
- Which categories you're interested in predicting.
- How many objects you want to use as reference for the prediction.
- For these objects, which features are relevant to the prediction.

This Learning set will then be used for running the prediction task.

Filter on title : # Matching features ≥ 10 Instrument : Zooscan × Filter

Project deep features model: LOKI_2022-05-17.
Model is usable.

	Inst.	# - Title	# Validated	# Matching features	Deep features model
☑	Zooscan	#859 - Demo Zooscan JO low taxo res	0	69	LOKI_2022-05-17
☐	Zooscan	#2374 - North Inlet Zooplankton	1893472	69	
☐	Zooscan	#1040 - _Zooscan_Reference	1444921	69	zooscan
☐	Zooscan	#6212 - ZOOSCAN PELGAS_2004_2016_NGrandremy_PhD_thesis	1153507	69	zooscan
☐	Zooscan	#6729 - ZOOSCAN PELGAS Pour Laetitia	477713	69	zooscan
☐	Zooscan	#2291 - TRAFFIC_M153_Multinet_Maxi	465089	69	zooscan
☐	Zooscan	#5421 - Gradients_v2	428099	69	zooscan
☐	Zooscan	#714 - Zooscan Tara Oceans MTN 300 ALL NETS - NEW!	393382	69	zooscan
☐	Zooscan	#6285 - Zooplankton Helgoland - Normalnetz 150 2006	382243	69	zooscan
☐	Zooscan	#1125 - ICaWR	350447	69	
☐	Zooscan	#4565 - zooscan_cecilia_2020	323013	69	
☐	Zooscan	#5977 - Zooscan_SO21_MN_0-1000m	313351	69	zooscan
☐	Zooscan	#611 - Projet DVM Lacs Boreaux - Sabrina Gignac Brassard 2016-2017	304151	69	
☐	Zooscan	#3657 - Zooscan Triatlas 2019 MTN 200	285458	69	zooscan
☐	Zooscan	#6111 - Triatlas 2019 MTN 200 NEW	285458	69	zooscan
☐	Zooscan	#418 - Zooscan Tara Oceans 2009 2013 MTN 300 sn033	284105	69	zooscan

How?

Step 3: Train a classifier based on classified images and predict the others

EcoTaxa 2.6 Demo Zooscan JO low taxo res

Jean-Olivier Irisson (log out)

Action

PREDICTION: Choice of Learning Set categories and size

Next: Choose features in selected objects

From data source, which is : #859 - Demo Zooscan JO low taxo res, only objects validated in below chosen categories will be present in the Learning Set being built.
Optionally, each category can **appear** as another category, generally a parent one, to the machine learning algorithm.

- The experience shows that it is often more efficient to predict into a limited number of categories and then validate in detail using more categories.

Learn from max. 5000 objects per category.
Total is currently 14 objects.

Make categories appear like in: Project search

(id)		Source (validated) category	# source	% source	# learning set	Appear as category
(84963)	☑	detritus	6	42.9 %	6	select category...
(25828)	☑	Copepoda	3	21.4 %	3	select category...
(85076)	☑	fiber < detritus	3	21.4 %	3	select category...
(85123)	☑	Appendicularia	2	14.3 %	2	select category...

How?

Step 3: Train a classifier based on classified images and predict the others

EcoTaxa 2.6 Demo Zooscan JO low taxo res

Jean-Olivier Irisson (log out)

Action

PREDICTION: Choice of features and settings

Start prediction task

Add deep features ☑

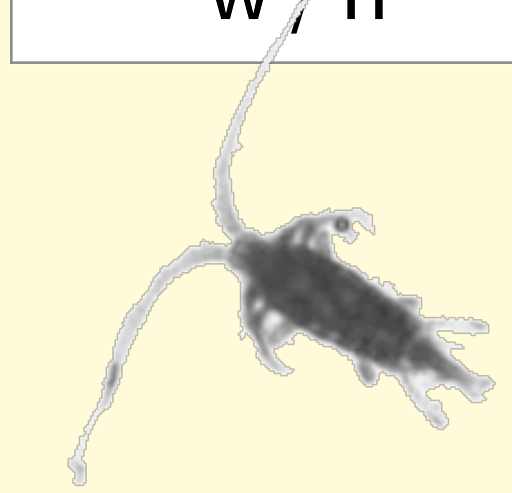
You have chosen 14 reference objects to build the Learning Set. In this last step, you can choose which features to associate with each of these objects, and start a prediction task using the Learning Set.

- Prediction will be better if you exclude features which are not related to the classification, e.g. coordinates in the raw image.
- Features with a single, constant value, or too many missing values, are useless for prediction and are automatically excluded. Some of them are listed here as a reminder.
- Missing values will be replaced by the median value for this feature from the reference objects.
- Prediction settings are recorded in EcoTaxa for the next prediction.

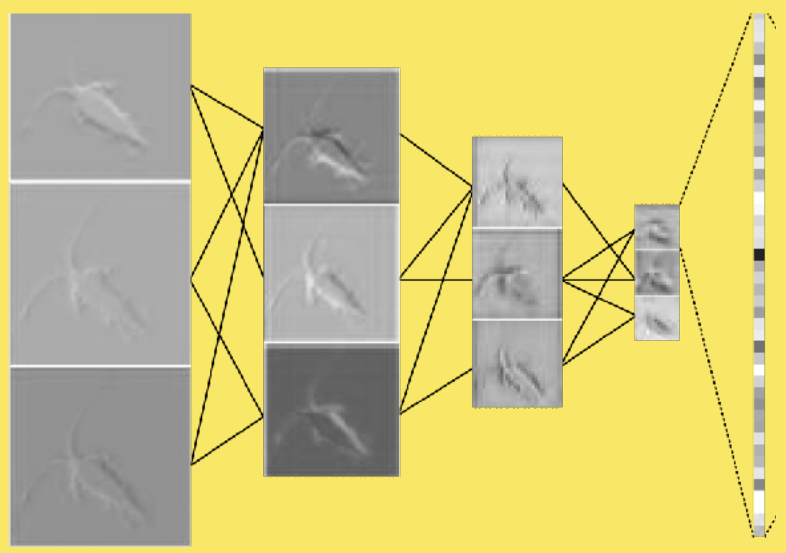
%area ☑	angle ☑	area ☑	area_exc ☑	bx ☑	by ☑	cdexc ☑	centroids ☑	circ. ☑	circex ☑	compentropy ☐	compm1 ☐
compm2 ☐	compm3 ☐	compmean ☐	compslope ☐	convarea ☑	convperim ☑	cv ☑	depth_max ☐	depth_min ☐	elongation ☑	esd ☑	fcons ☑
feret ☑	feretareaexc ☑	fractal ☑	height ☑	histcum1 ☑	histcum2 ☑	histcum3 ☑	intden ☑	kurt ☑	lat_end ☐	lon_end ☐	major ☑
max ☑	mean ☑	meanpos ☑	median ☑	min ☑	minor ☑	mode ☑	nb1 ☑	nb2 ☑	nb3 ☐	perim. ☑	perimareaexc ☑
perimferet ☑	perimmajor ☑	range ☑	skelarea ☑	skew ☑	slope ☑	sr ☑	stddev ☑	symetrieh ☑	symetriehc ☑	symetriev ☑	symetrievc ☑
tag ☐	thickr ☑	width ☑	x ☑	xm ☑	xmg5 ☐	xstart ☑	y ☑	ym ☑	ymg5 ☐	ystart ☑	

A note on the machine learning architecture

IMAGE
path
rank
w / h



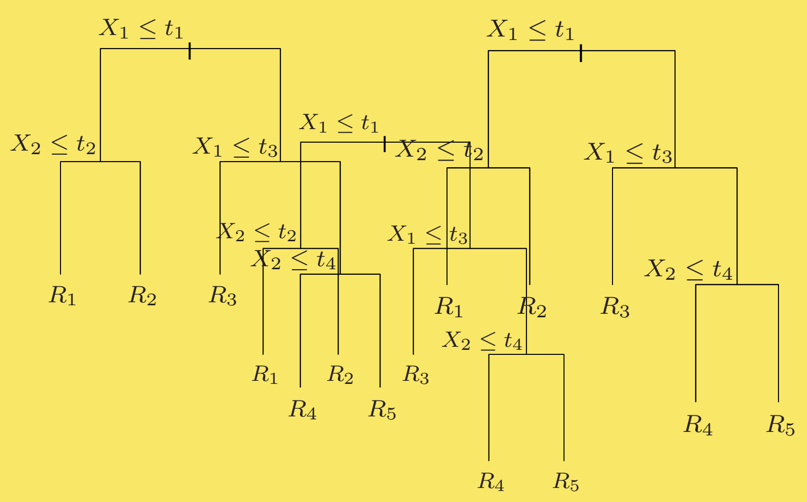
CNN



OBJECT
objID
taxoID
area
major ax.
minor ax.
feature 4
...
feature n

CNN
CNN feature 1
...
CNN feature n

Random forest



taxoID

score

How?

Step 3: Train a classifier based on classified images and predict the others

EcoTaxa 2.6

Pending : 1 Error :1

Jean-Olivier Irisson (log out)

Action

25% Retrieving objects to classify

WARNING : this is an asynchronous process.

You can either:

- Open another tab or window to continue working in EcoTaxa (and monitor the progress of the task switching to this page from time to time)
- Leave this page to continue working in EcoTaxa (and click on the "Run" or "Done" buttons on the upper right of the Ecotaxa topper to monitor the progress of the task or check its completion)

Asynchronous tasks are utilized for processes that can be long. They can create temporary data (like export file) that can be manually removed at the end of the process even if EcoTaxa will do it for you on a regular schedule. Remaining tasks can be checked using the "Done" or "Error" top right buttons.

How?

Step 4: Sort predicted images by classification score

How?

Step 4: Sort predicted images by classification score

How?

Step 5: Confirm or correct automatic classifications

How?

Step 5: Confirm or correct automatic classifications

How?

Step 5: Confirm or correct automatic classifications

How?

Rinse and repeat!

How?

Rinse and repeat!

Eco**Taxa** today

242M objects

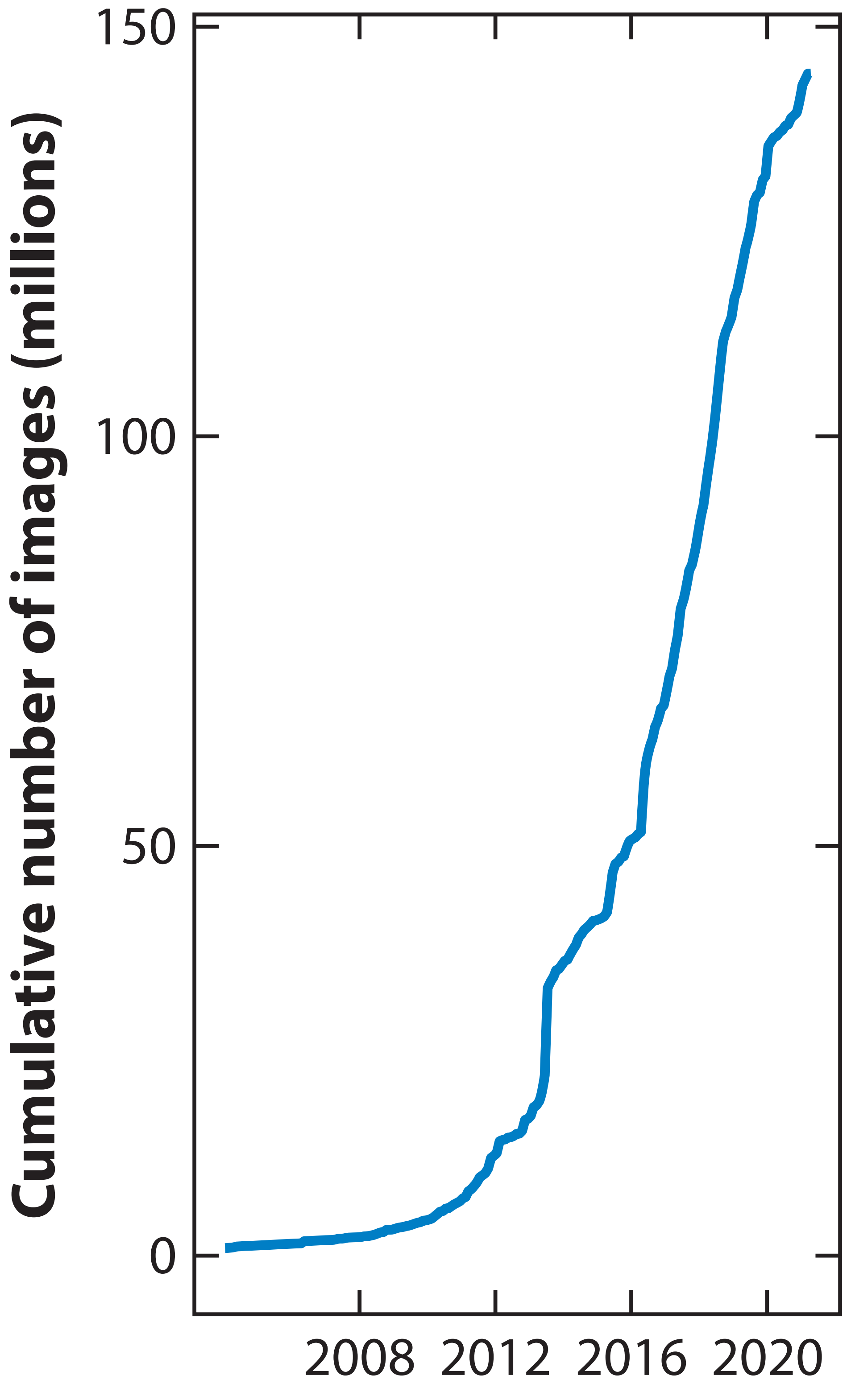
101M classified

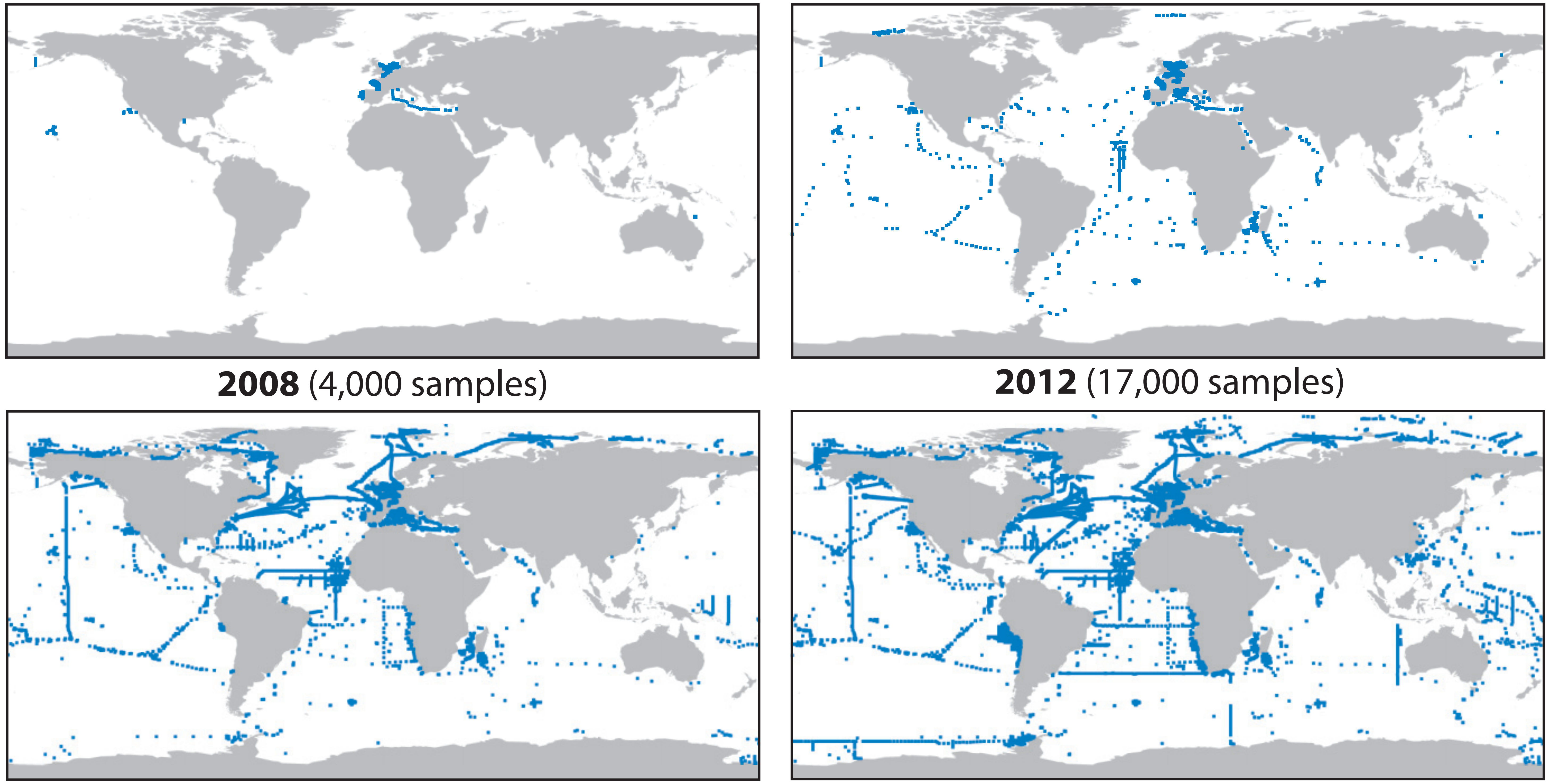
+5M per month

1,800 users from **550 institutions** worldwide

At any point of the day, 90 sessions from **50 users**

Sorting of **5,000 to 10,000 images** per day per operator



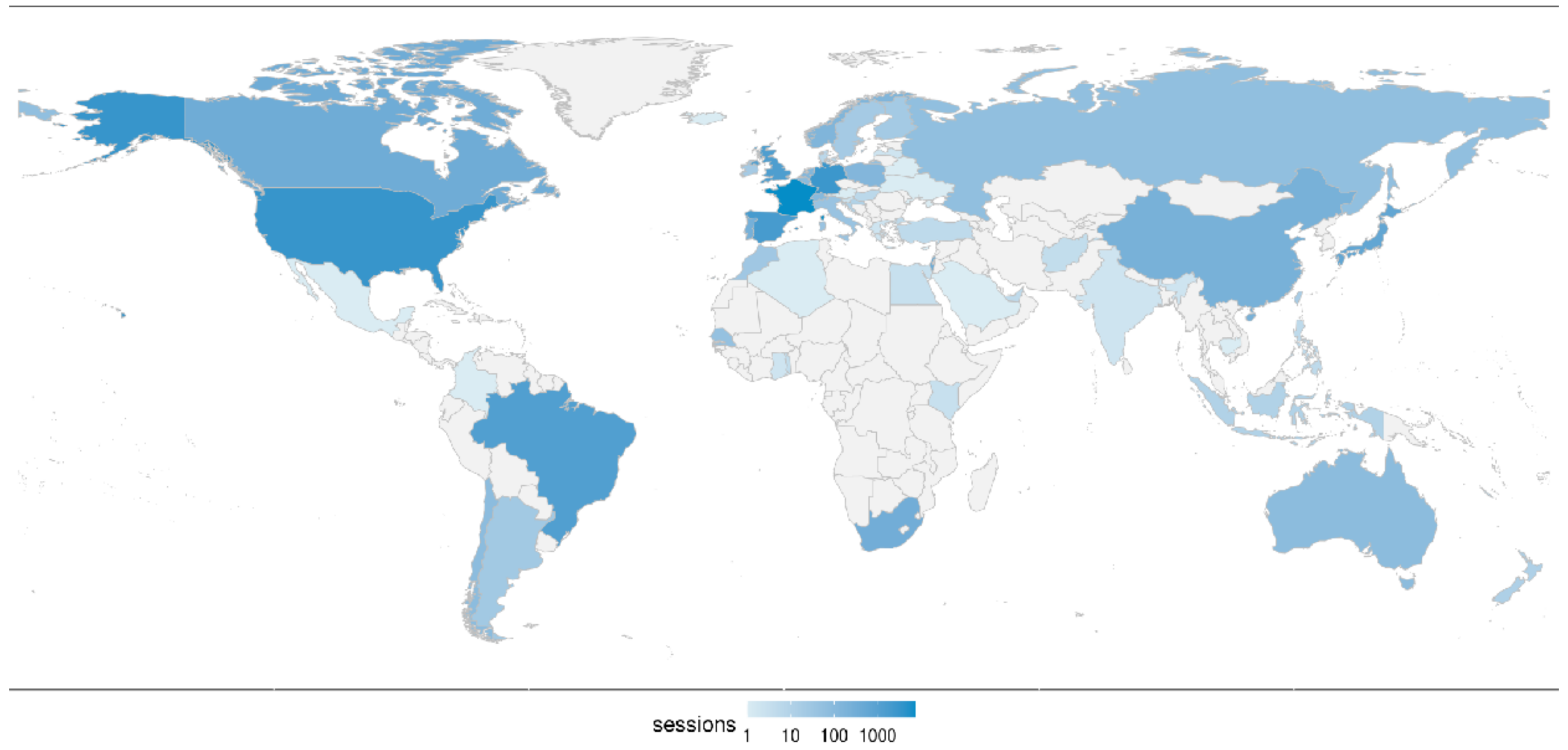


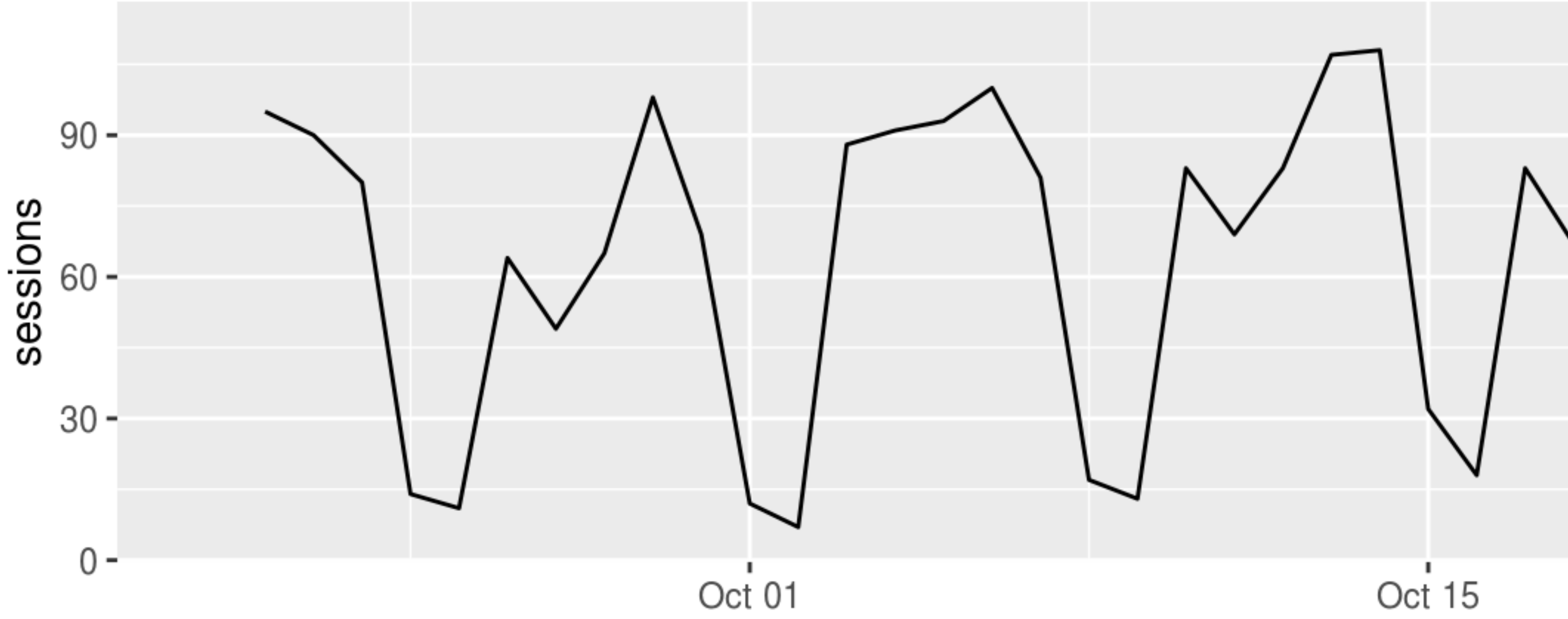
2008 (4,000 samples)

2012 (17,000 samples)

2016 (56,000 samples)

2020 (91,000 samples)





How deep is enough?

Model	Size	Accuracy	Avg. precision	Avg. recall
Classic features + Random Forest	1M	71.1	65.0	64.2

How deep is enough?

Model	Size	Accuracy	Avg. precision	Avg. recall
Classic features + Random Forest	1M	71.1	65.0	64.2
MobileNet v4 + 600	5.4M	89.4	91.2	92.0

How deep is enough?

Model	Size	Accuracy	Avg. precision	Avg. recall
Classic features + Random Forest	1M	71.1	65.0	64.2
MobileNet v4 + 600	5.4M	89.4	91.2	92.0
MobileNet v4 + 1792	7.5M	89.2	90.9	91.9

How deep is enough?

Model	Size	Accuracy	Avg. precision	Avg. recall
Classic features + Random Forest	1M	71.1	65.0	64.2
MobileNet v4 + 600	5.4M	89.4	91.2	92.0
MobileNet v4 + 1792	7.5M	89.2	90.9	91.9
EfficientNet v2 S + 600	25M	89.8	91.2	92.9

How deep is enough?

Model	Size	Accuracy	Avg. precision	Avg. recall
Classic features + Random Forest	1M	71.1	65.0	64.2
MobileNet v4 + 600	5.4M	89.4	91.2	92.0
MobileNet v4 + 1792	7.5M	89.2	90.9	91.9
EfficientNet v2 S + 600	25M	89.8	91.2	92.9
EfficientNet v2 XL + 600	208M	89.1	90.9	92.3

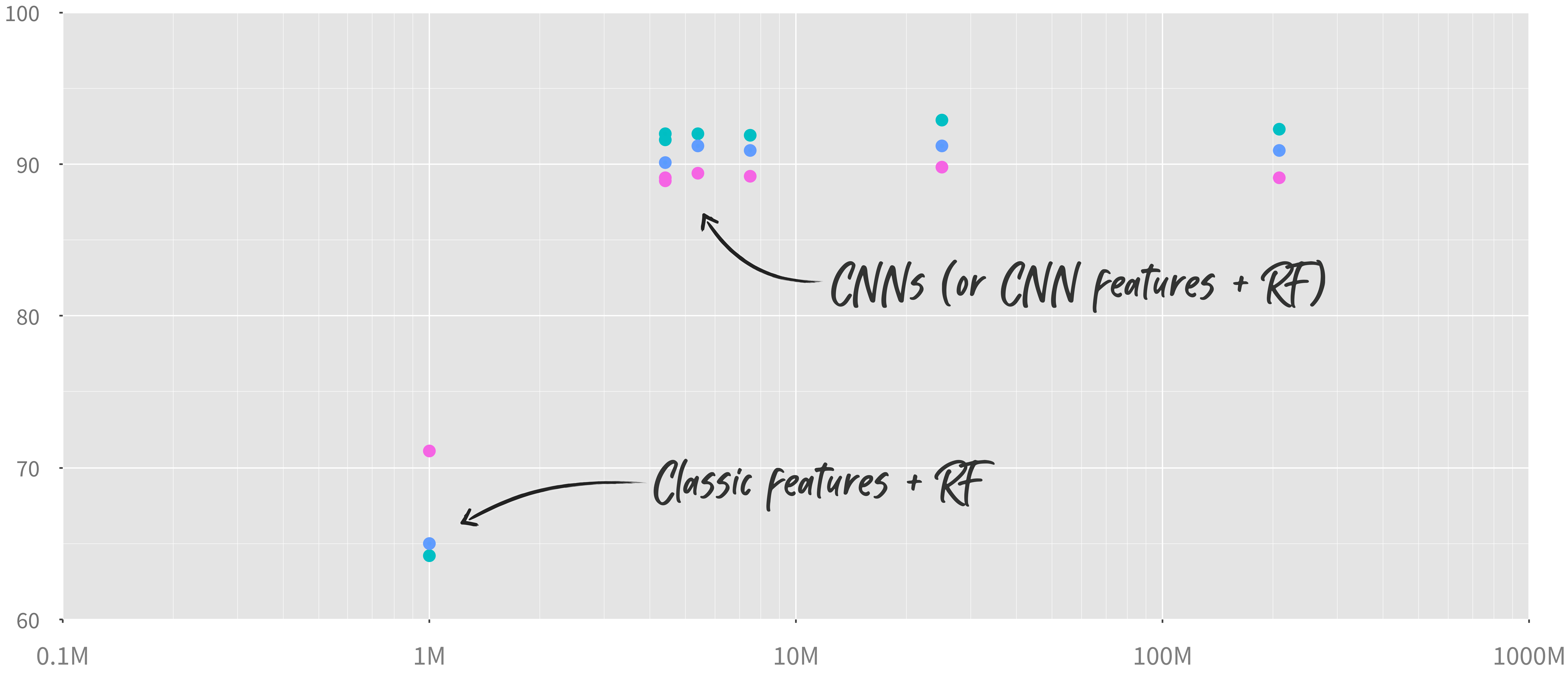
How deep is enough?

Model	Size	Accuracy	Avg. precision	Avg. recall
Classic features + Random Forest	1M	71.1	65.0	64.2
MobileNet v4 + 600	5.4M	89.4	91.2	92.0
MobileNet v4 + 1792	7.5M	89.2	90.9	91.9
EfficientNet v2 S + 600	25M	89.8	91.2	92.9
EfficientNet v2 XL + 600	208M	89.1	90.9	92.3
MobileNet v4 + 50	4.4M	88.9	90.1	901.6

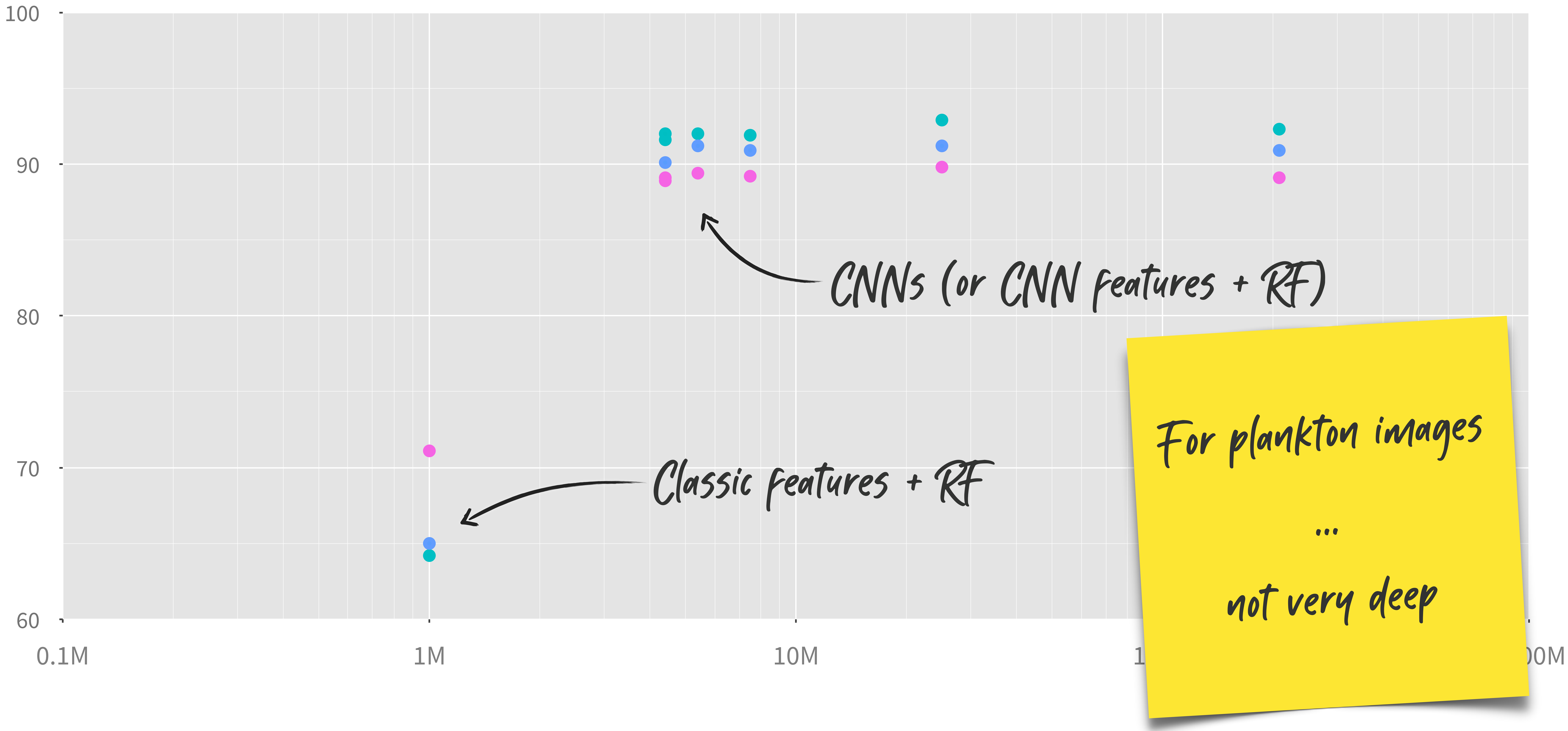
How deep is enough?

Model	Size	Accuracy	Avg. precision	Avg. recall
Classic features + Random Forest	1M	71.1	65.0	64.2
MobileNet v4 + 600	5.4M	89.4	91.2	92.0
MobileNet v4 + 1792	7.5M	89.2	90.9	91.9
EfficientNet v2 S + 600	25M	89.8	91.2	92.9
EfficientNet v2 XL + 600	208M	89.1	90.9	92.3
MobileNet v4 + 50	4.4M	88.9	90.1	901.6
MobileNet v4 + 1792 + PCA 50 + RF	~4.4M	89.1	90.1	92.0

How deep is enough?



How deep is enough?

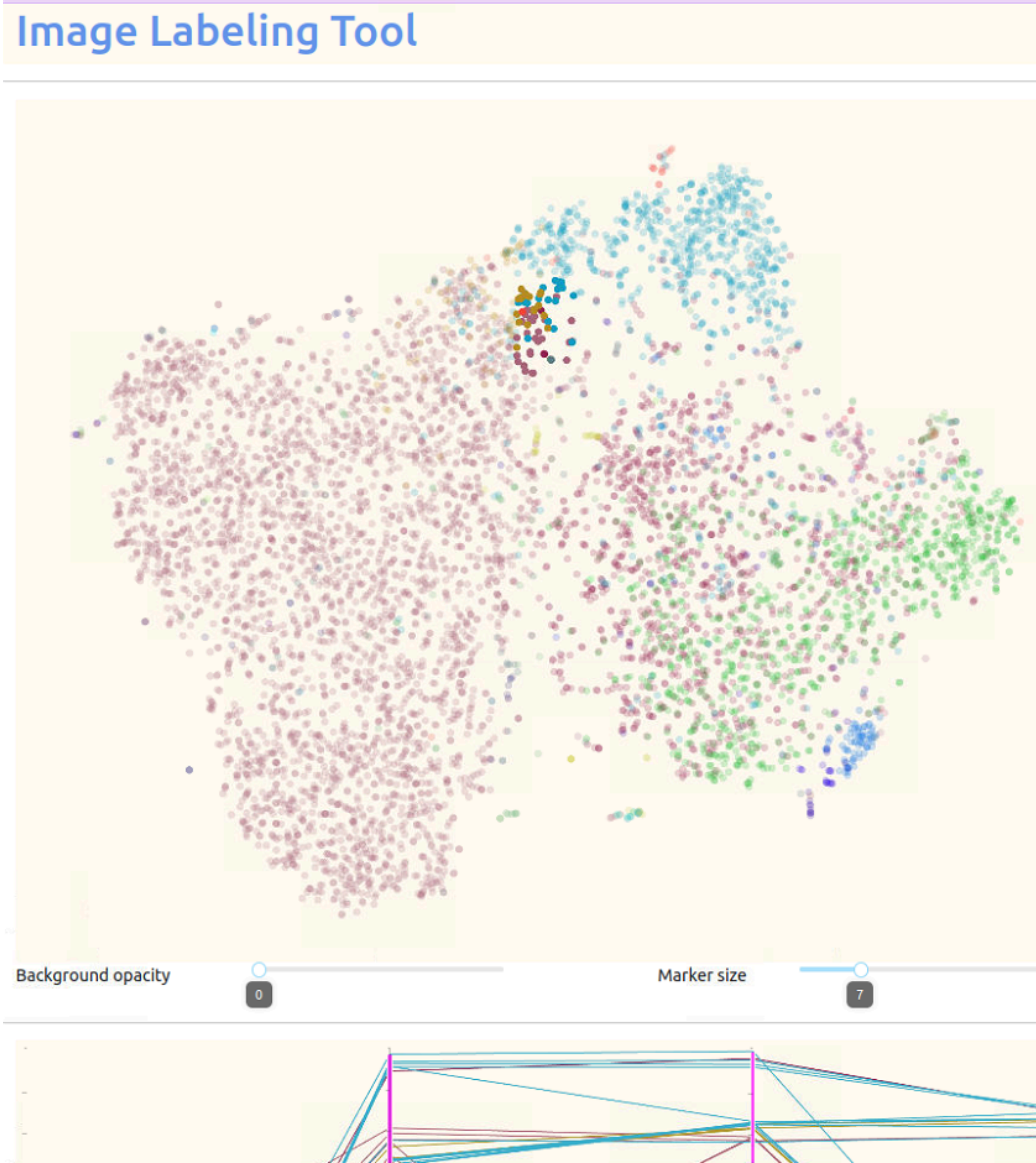


Future plans

Search by **similarity**/dissimilarity

Batch classification

Integration with a dedicated **taxonomic guide**



Future plans

Search by **similarity**/dissimilarity

Batch classification

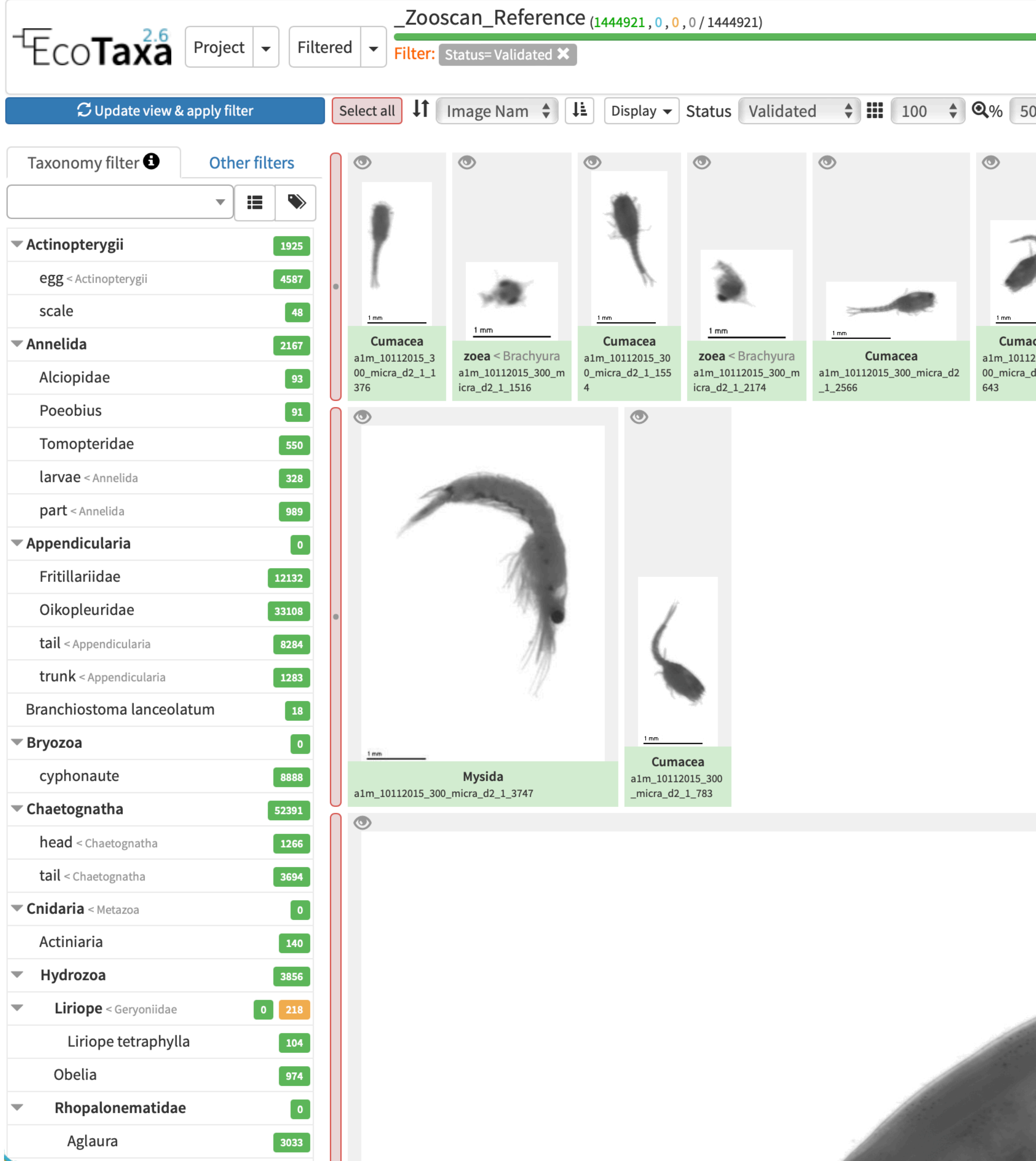
Integration with a dedicated **taxonomic guide**

Future plans

Search by **similarity**/dissimilarity

Batch classification

Integration with a dedicated **taxonomic guide**



EcoTaxa specifi...

2.5 Sharing reference list of categories

The updated list of categories must be portable in order to be able to update an application running on a field computer without any Internet connexion.

2.6 Quality flag

The **quality** of the identification is noted by a **flag** (*annotation_status*) which is automatically set by the application in most cases.

The quality flag can be read at the importation stage if exists in the spreadsheet provided with the images.

The possible values are:

- No identification ("noid")
- Identification from the automated prediction output ("predicted")
- A manual doubtful identification needs to be confirmed by other annotators ("doubtful")
- Identification is fully validated ("validated")

- Flag changes from "predicted" to "validated" when the vignette is manually dragged into a category or renamed or when the "validate all" button is pressed.
- Flag changes to "doubtful" when selected in the object page.
- Flag changes from "noid" to "predicted" when an automatic prediction is made
- Flag changes from "noid" to "validated" when the vignette is manually dragged into a category (ie when it is manually and intentionally renamed).

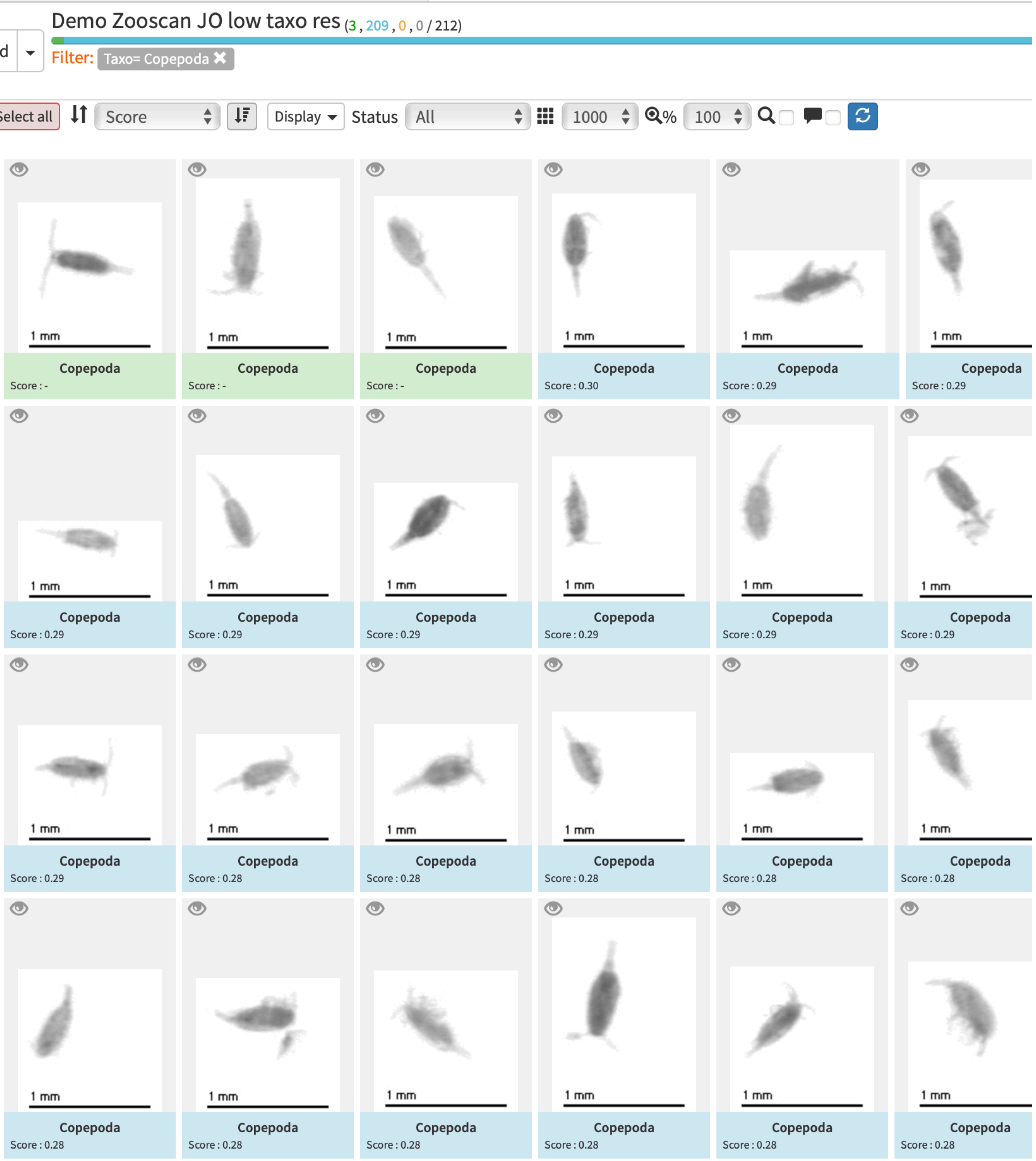
2.7 Users roles and rights

Four roles of users and associated rights/privileges are considered:

	visitor	annotator	project manager	application manager (unique)
View images	X	X	X	X
Annotate images (manual)		X	X	X
Register annotators			X	X
Annotate images (automatic)			X	X
Import new objects in project			X	X
Set project status			X	X
Create subset from project			X	X
Manage the taxonomy list			X	X
Create classes			X	X
Create project				X
Register project managers				X
		X (with notification		

Lessons learned

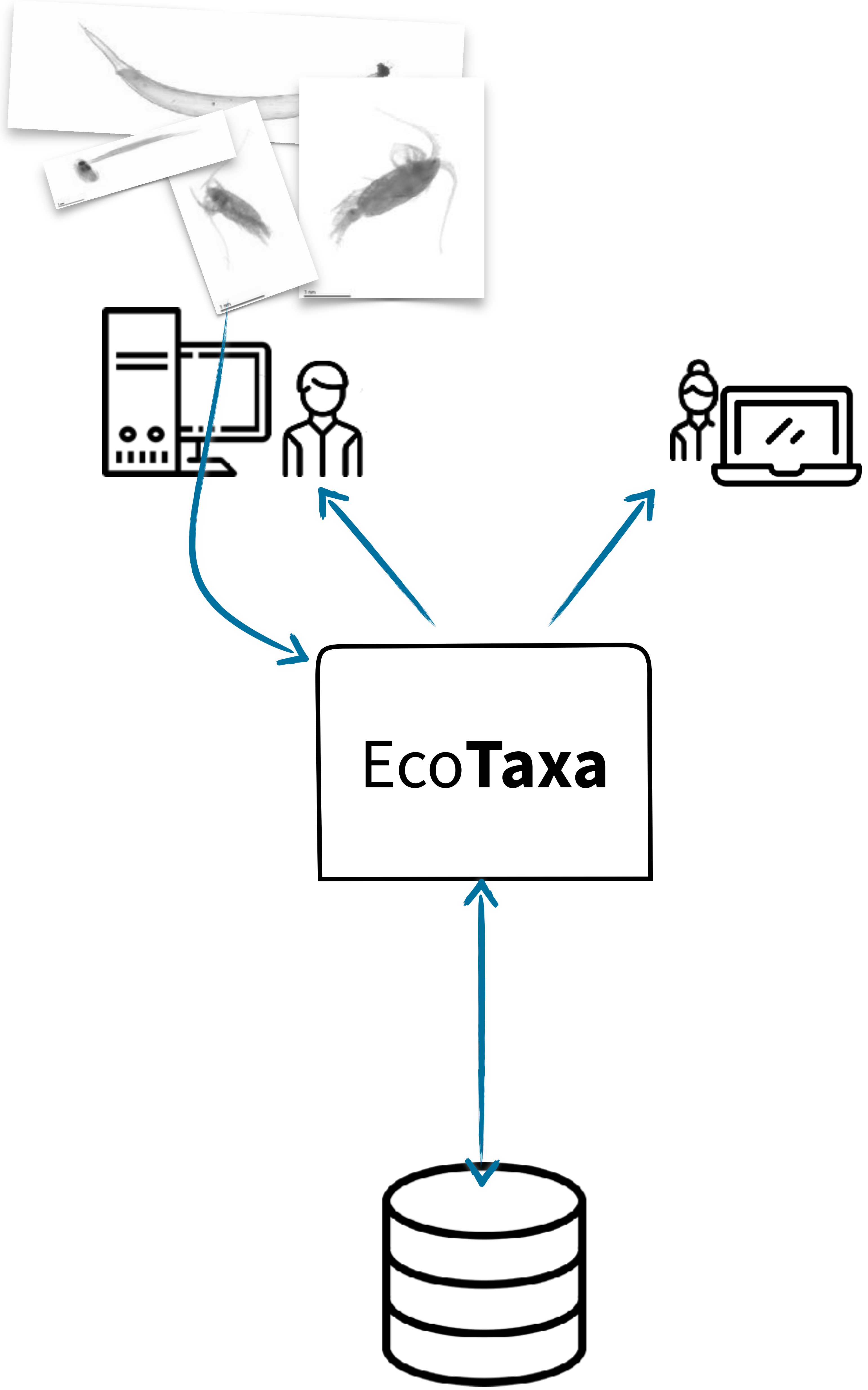
Subcontracting has the advantage of forcing to **specify** features



Lessons learned

Subcontracting has the advantage of forcing to **specify** features

Machine learning as an **assistance** rather than an end in itself; and **KISS**

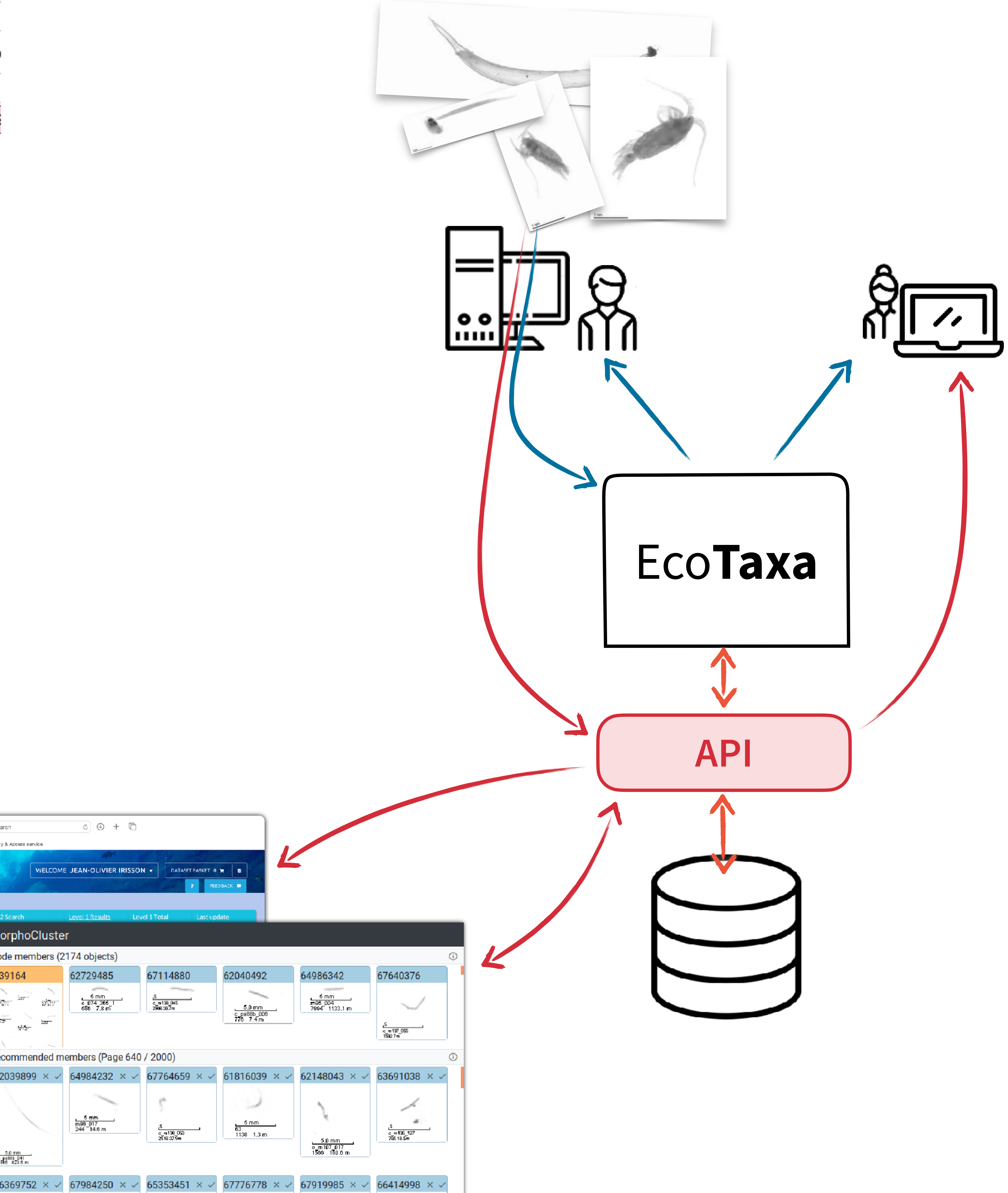


Lessons learned

Subcontracting has the advantage of forcing to **specify** features

Machine learning as an **assistance** rather than an end in itself; and **KISS**

An **Application Programming Interface** is the center of the modern web

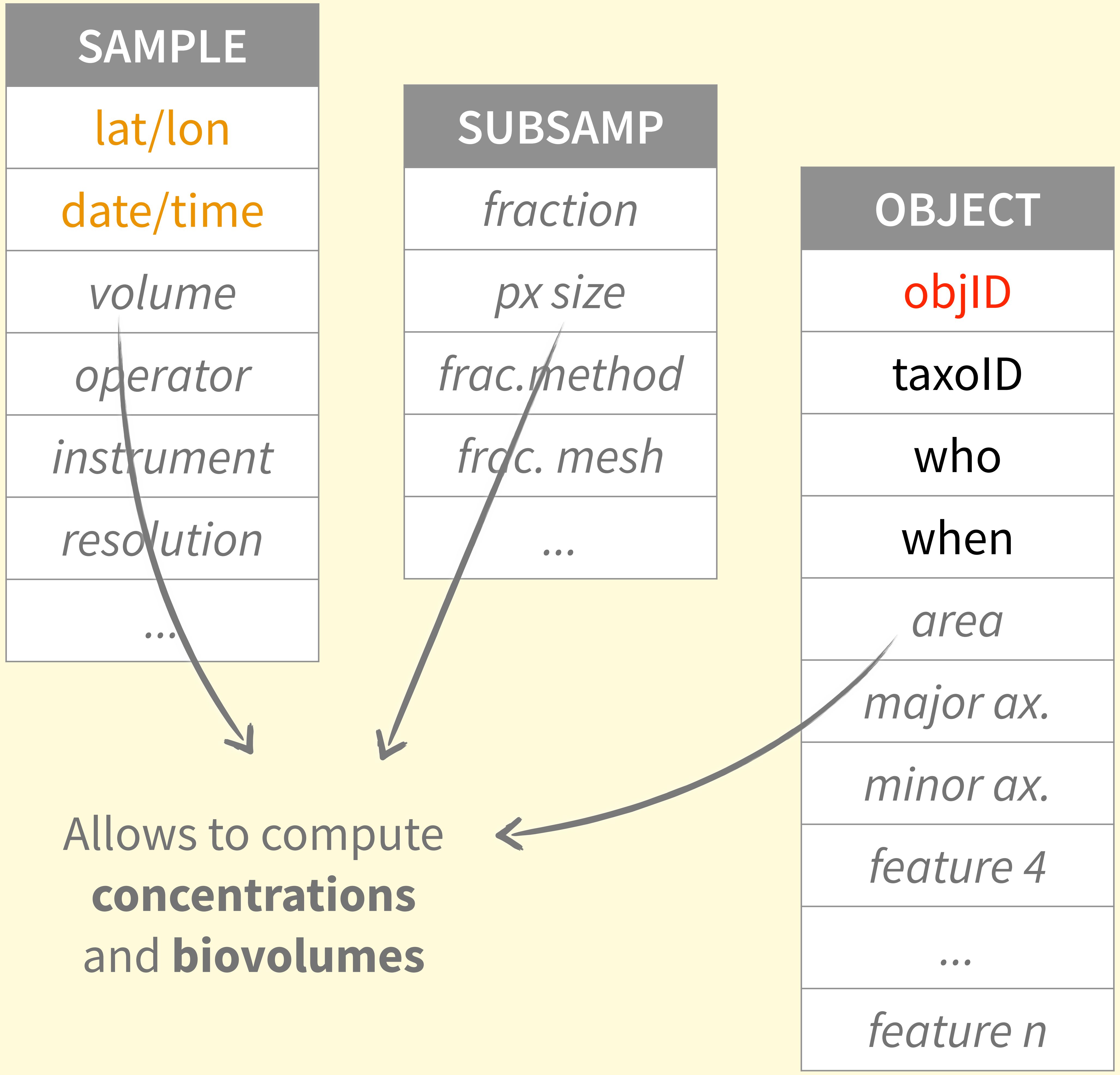


Lessons learned

Subcontracting has the advantage of forcing to **specify** features

Machine learning as an **assistance** rather than an end in itself; and **KISS**

An **Application Programming Interface** is the center of the modern web



Lessons learned

Subcontracting has the advantage of forcing to **specify** features

Machine learning as an **assistance** rather than an end in itself; and **KISS**

An **Application Programming Interface** is the center of the modern web

Flexibility is good, too much flexibility is **difficult**

Lessons learned

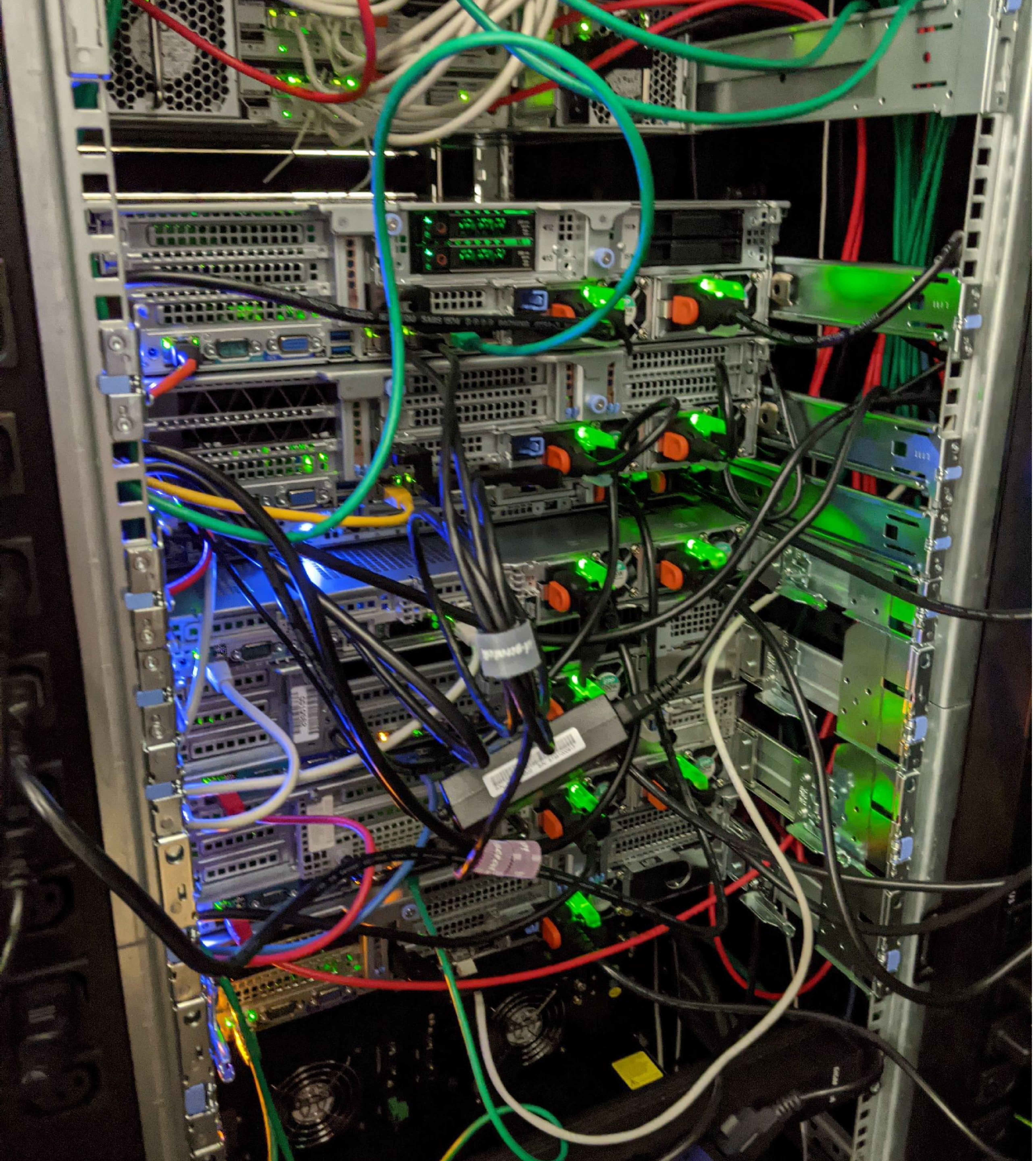
Subcontracting has the advantage of forcing to **specify** features

Machine learning as an **assistance** rather than an end in itself; and **KISS**

An **Application Programming Interface** is the center of the modern web

Flexibility is good, too much flexibility is **difficult**

Proper (web) development is **expensive**



Lessons learned

Subcontracting has the advantage of forcing to **specify** features

Machine learning as an **assistance** rather than an end in itself; and **KISS**

An **Application Programming Interface** is the center of the modern web

Flexibility is good, too much flexibility is **difficult**

Proper (web) development is **expensive**

Institutional **hosting** is a problem

cnrs

SCIENCES SORBONNE UNIVERSITÉ

European Commission HORIZON 2020

FACE FOUNDATION

BELMONT FORUM