MEGALODOM, 2017-06-22
Jean-Olivier Irisson

Machine learning for the classification of plankton images

Overcoming the data scarcity… and then overflow

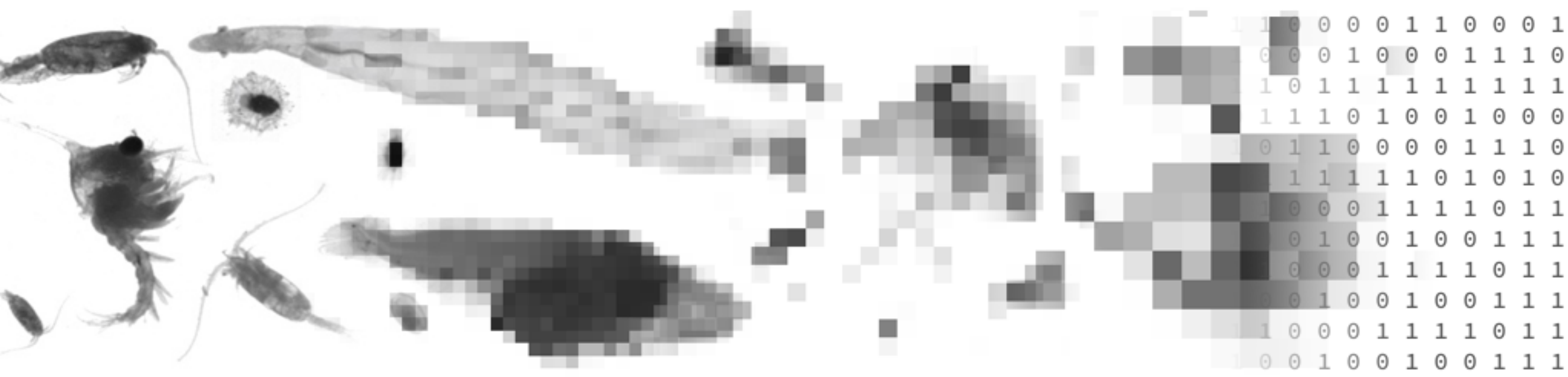



Image Landsat / Copernicus
Image IBCAO
Data SIO, NOAA, U.S. Navy, NGA, GEBCO


The problem
DATA

One solution

Pros

High **taxonomic** resolution

Cons

Requires a lot of **time** (of experts)

Only **abundance** information

Not easily **replicable** (human error scarcely evaluated)

Another solution

Digitization

Segmentation

Classification

Another solution

Digitization

Segmentation

Classification

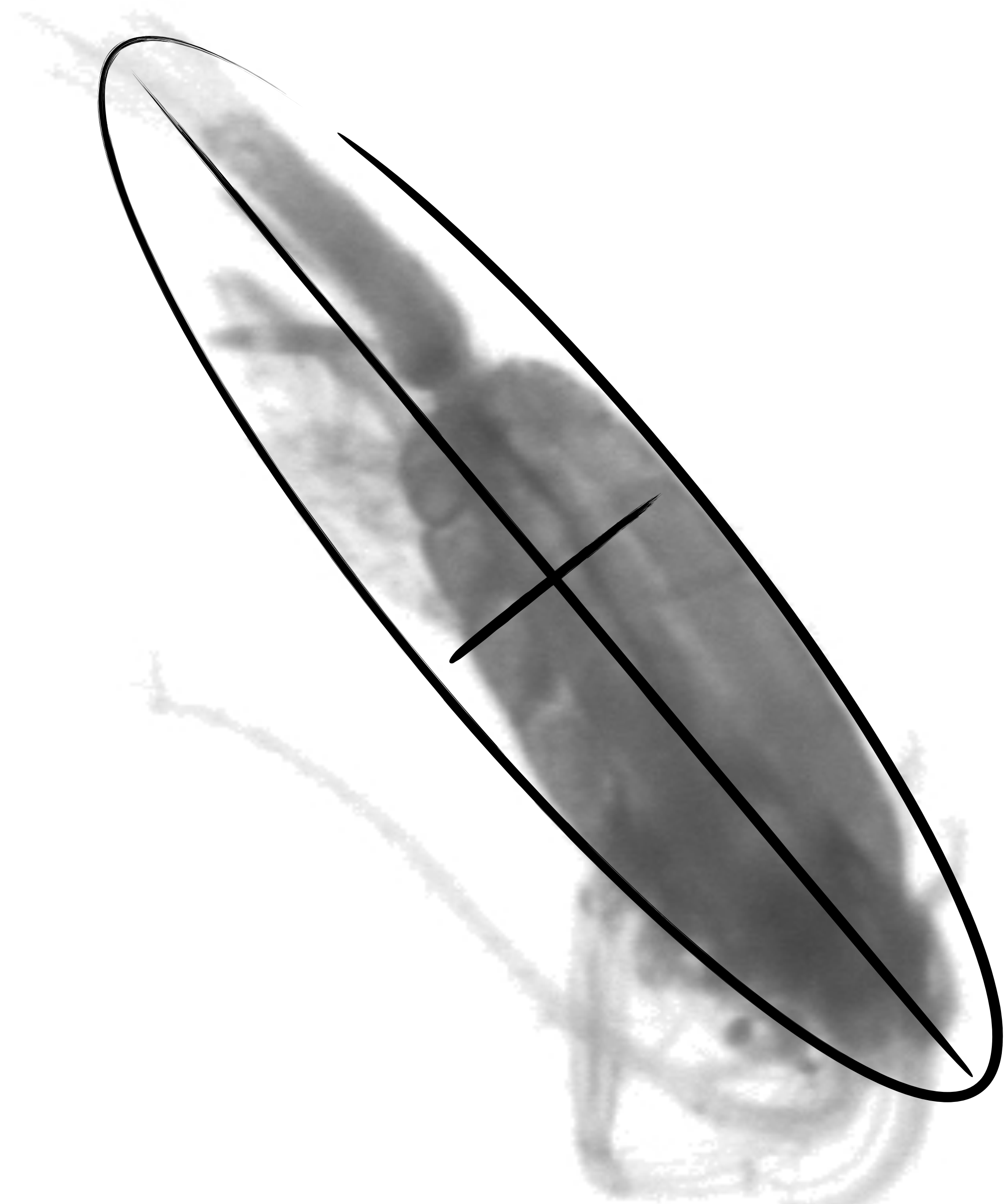
Another solution

Digitization

Segmentation

Classification

Another solution



Digitization

Segmentation

Classification

nm

Various classifiers

Since 2004 (on a Sun SPARCstation 20!)

RandomForest

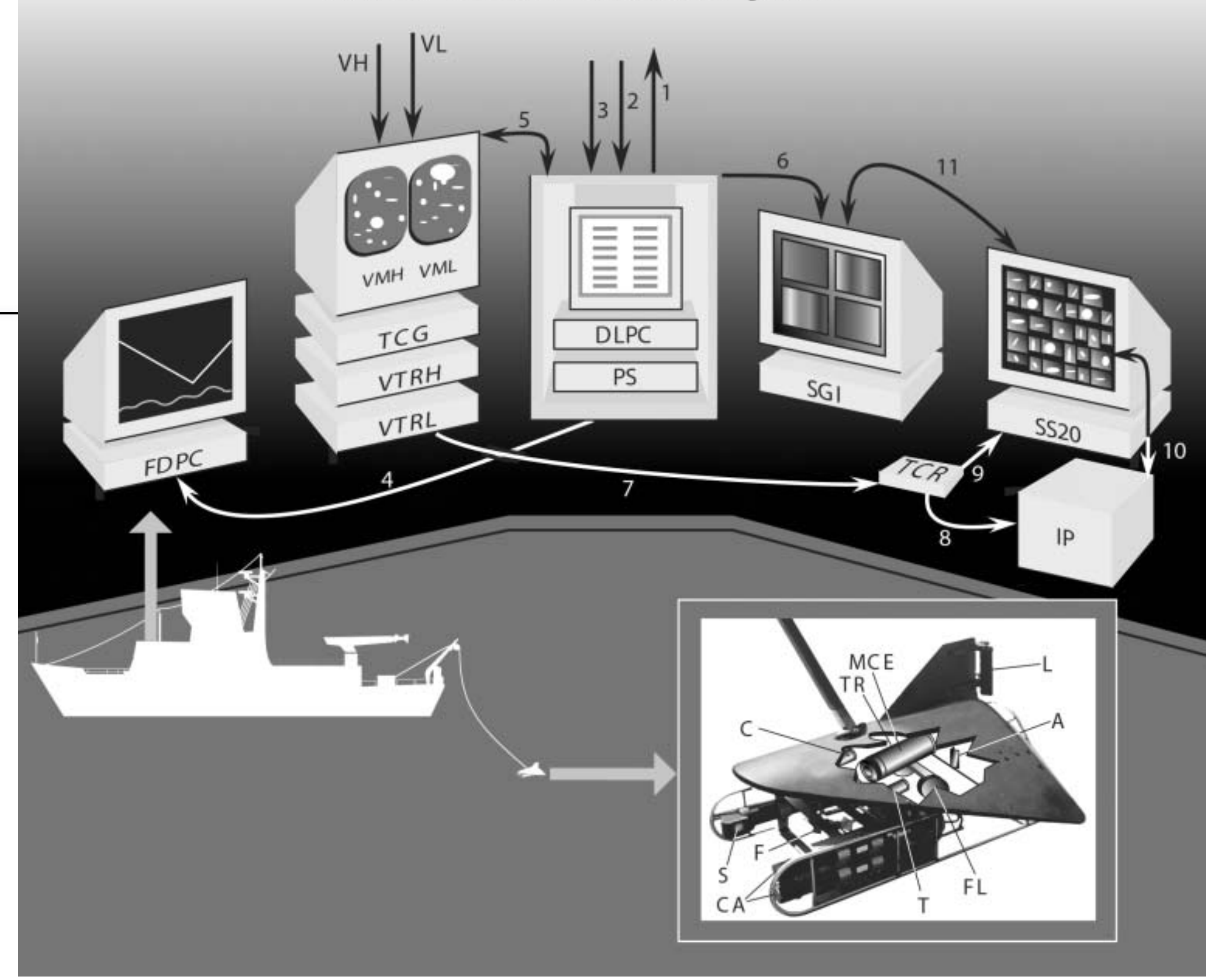
Support Vector Machines

Naïve Bayesian Classifier

Various neural networks

...

+ combination of the above



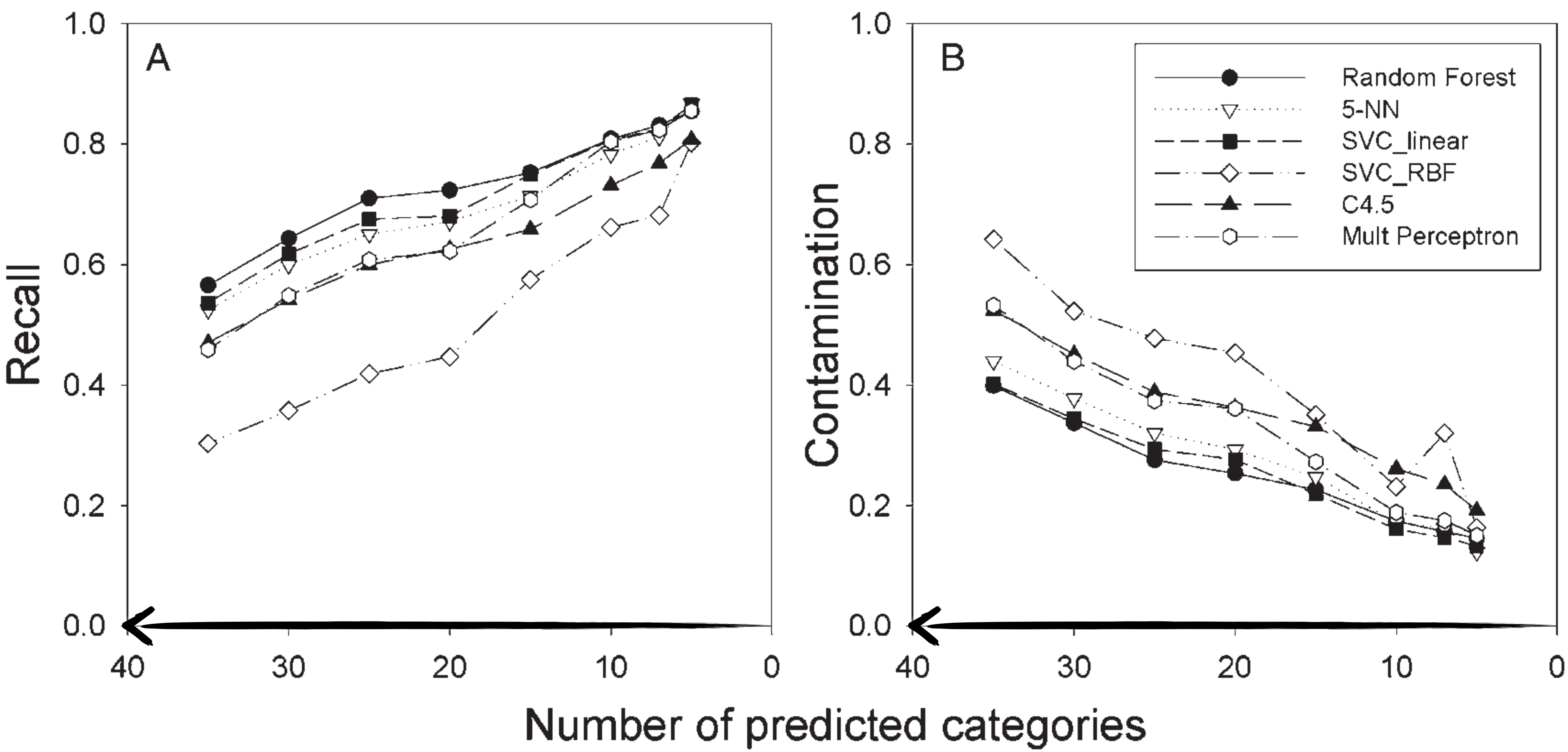
Imperfect automatic classification

2004

Method	Simplified (8 groups)		Detailed (29 groups)	
	Accuracy (%)	Speed (s)	Accuracy (%)	Speed (s)
Linear discriminant analysis	76.8	0.1	70.6	0.2
Quadratic discriminant analysis	82.9	0.2	–	–
Mixture discriminant analysis	81.4	2.4	–	–
Flexible discriminant analysis	77.6	1.8	72.7	6.0
k-nearest neighbour analysis	77.2	0.1	60.4	0.1
Learning vector quantization	76.6	0.3	60.0	0.4
Tree method	72.0	0.5	55.1	2.3
Recursive partitioning	72.8	1.2	57.7	3.1
Bagging (bootstrap on trees)	81.7	3.6	69.8	8.0
Double bagging with LDA	85.0	10.3	74.6	25.5
Double bagging with k-n.n.	81.9	8.9	70.1	13.8
Random forest	83.9	1.7	73.4	2.5
Support vector machine	68.5	1.2	47.8	1.9
Neural network	73.9	25.8	–	–
Discriminant vector forest	83.6	2.7	74.4	4.0

Imperfect automatic classification

2010



Our solution

+

Our solution

Pros

Mostly **replicable** and quite easy to evaluate

Provides information about **size**, transparency, etc. = functions

Can be done ***in situ***

Cons

Taxonomic resolution from automatic classification too **low** for many ecological studies

Still requires human **time** (not necessarily much faster)

Our solution

Pros

Mostly **replicable** and quite easy to evaluate

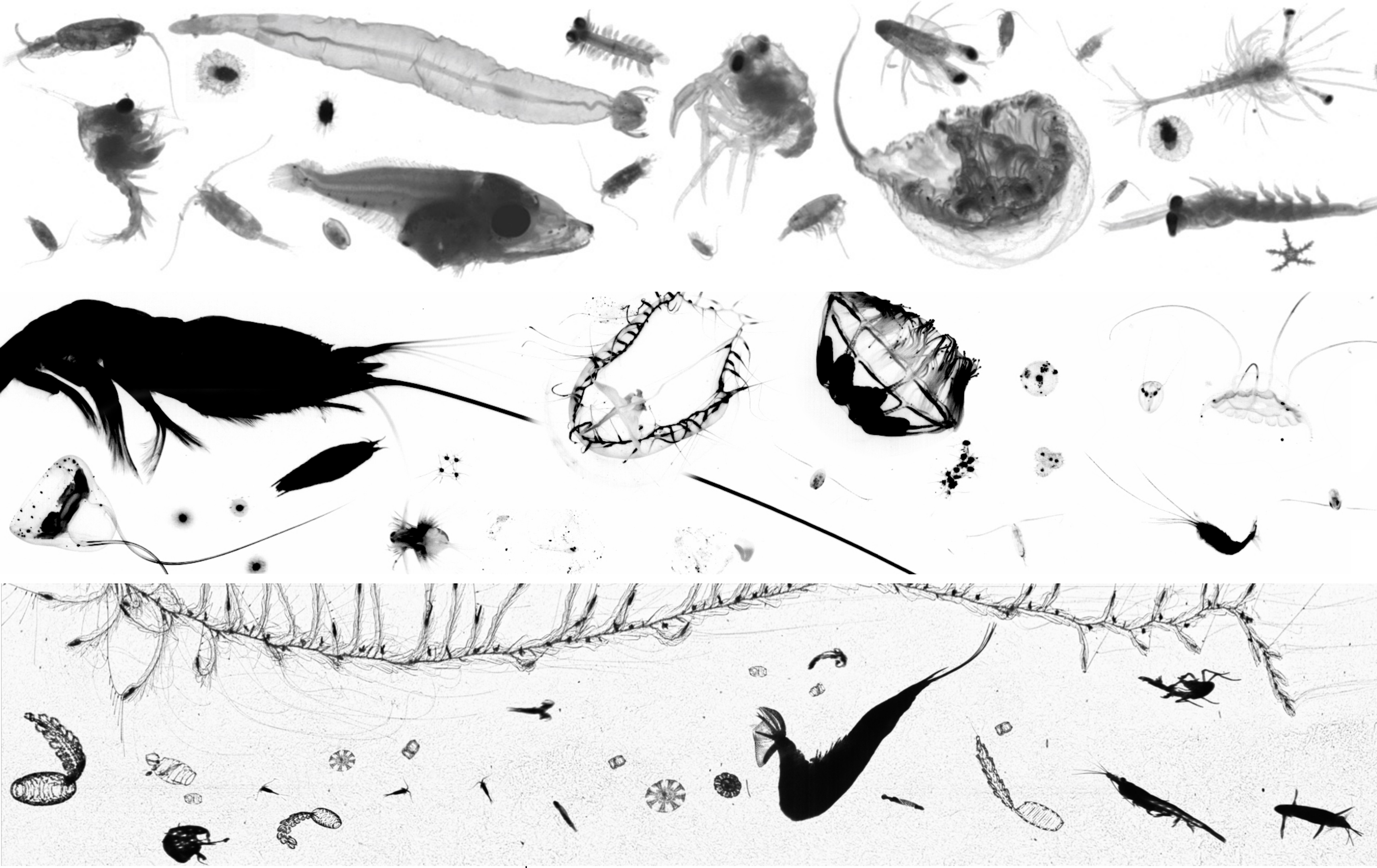
Provides information about **size**, transparency, etc. = functions

Can be done ***in situ***

Cons

Taxonomic resolution from automatic classification too **low** for many ecological studies

Still requires human **time** (not necessarily much faster)



Current flow of images

ZooScan = 1 Bpx/y, UVP = 8.6Bpx/y, ISIIS=25Tpx/y

⟹ Several million objects to classify per year

Kaggle 2015 competition

International competition for the classification of **plankton** images

60k images to classify in ~**120 groups** from a training set of 30k

1049 teams for a prize of $150k

Top 10 teams all used **CNNs**

83 to **85%** accuracy

SparseConvNet in 3rd place

www.kaggle.com/c/datasciencebowl

NATIONAL DATA SCIENCE BOWL

Completed • $175,000 • 1,049 teams

National Data Science Bowl

Mon 15 Dec 2014 – Mon 16 Mar 2015 (22 months ago)

Dashboard

- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
  - About the NDSB
  - Timeline
  - Tutorial
- Forum
- Leaderboard
  - Public
  - Private
- My Team
  - GitHub
- My Submissions

Private Leaderboard

1. ≈ Deep Sea ≈
2. Happy Lantern Festival
3. Poisson Process
4. Junonia
5. ⚓ Deepsea Challenger ⚓
6. AuroraXie
7. Maxim Milakov
8. Ilya Kostrikov
9. old-ufo
10. nagadomi

Forum (154 topics)

- scikit-learn Random Forest memory problem — 3 months ago
- Install Theano on Windows 8.1 with GPU enabled: pycuda installation problems — 4 months ago
- caffe training curves — 5 months ago
- Does anyone use caffee? How could I produce a test result? — 9 months ago
- Caffe? How to generate the prediction from caffe output? — 10 months ago
- Can someone explain what batch size is doing in convolutional NNs? — 13 months ago

Competition Details » Get the Data » Make a submission

Predict ocean health, one plankton at a time

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

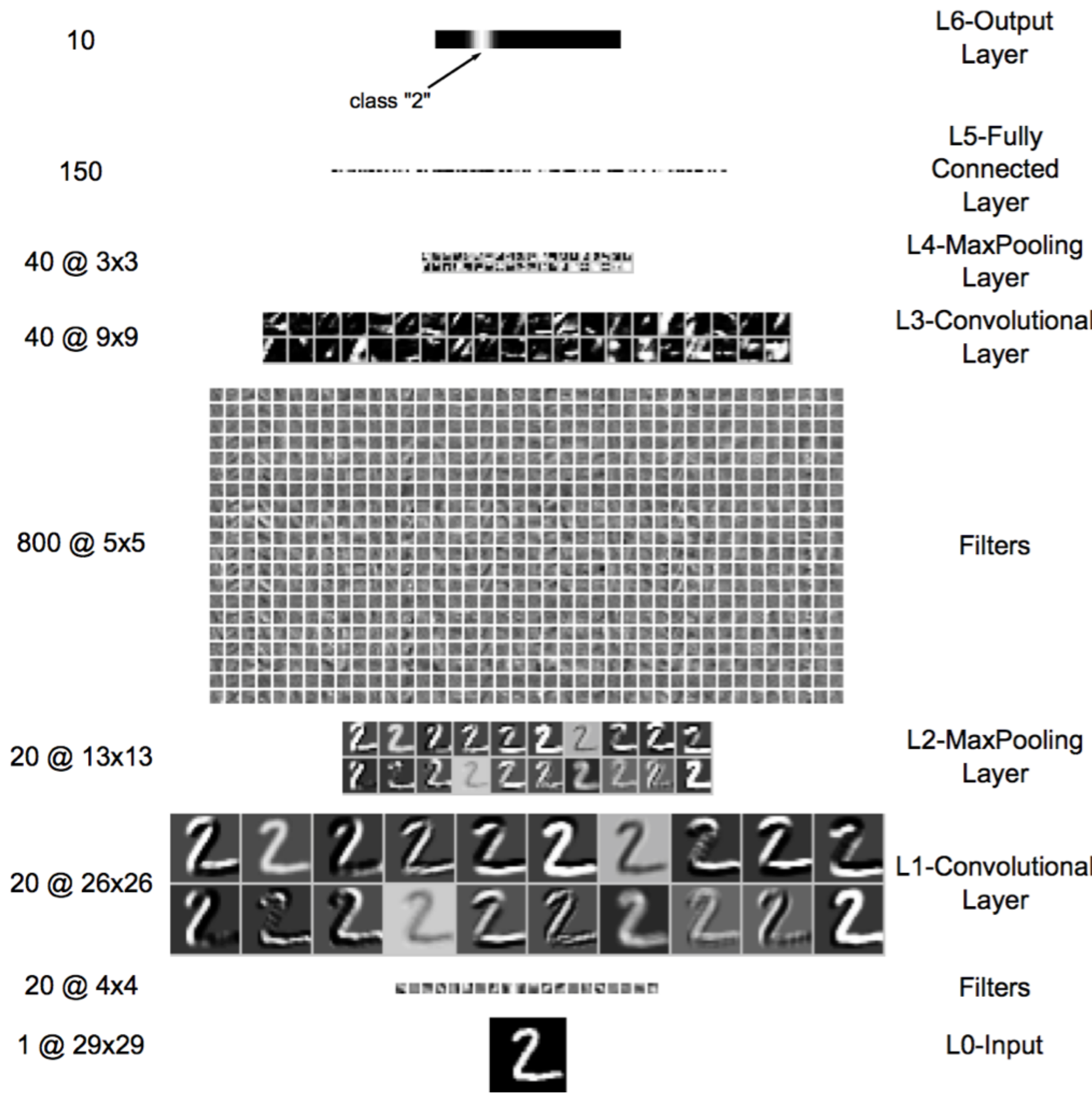
CNNs and SparseConvNet

https://github.com/btgraham/SparseConvNet

All custom C++

Sparsity

Fractional Max-Pooling



SparseConvNet + Zooscan

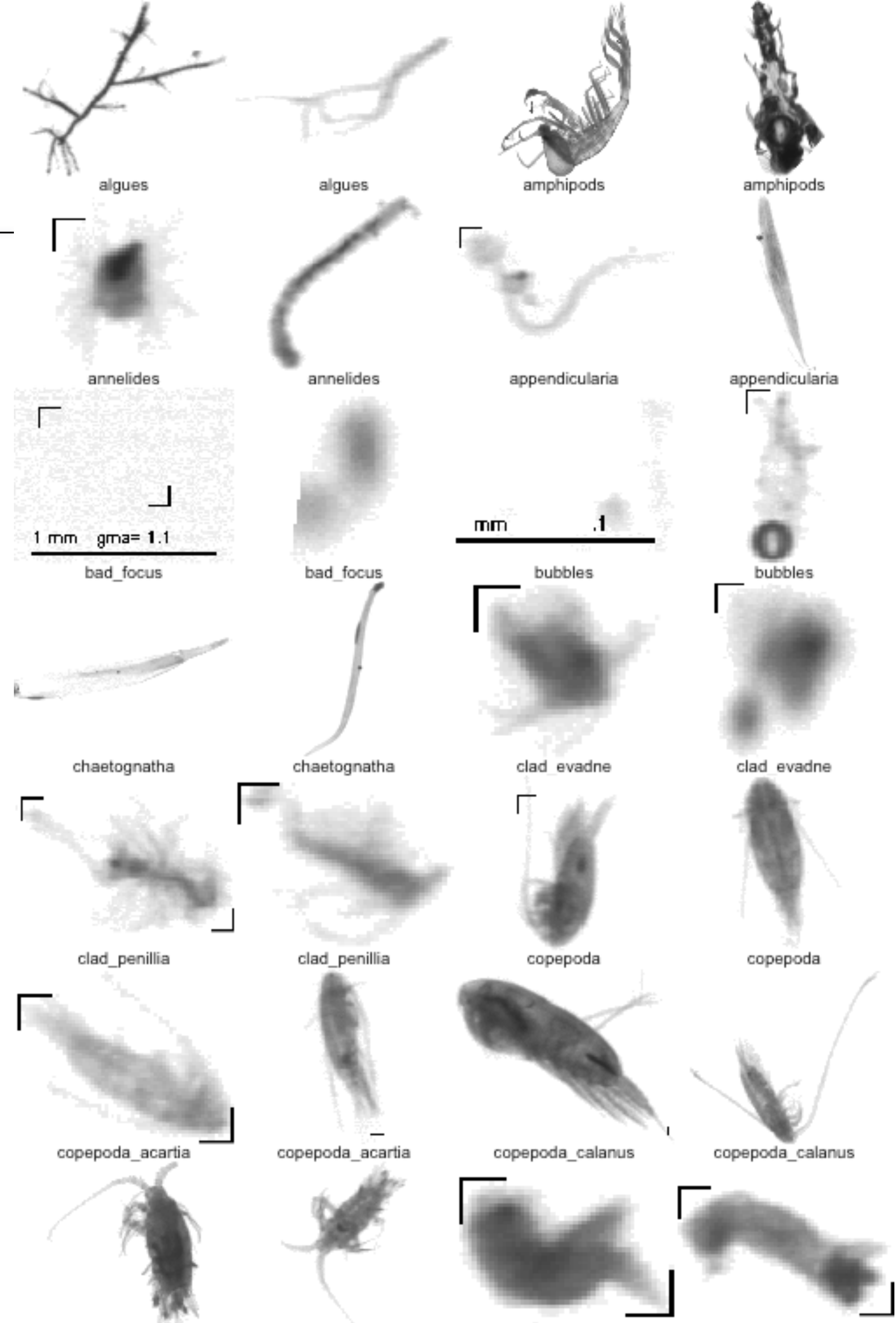
10k images for training, 80k for testing

Zooprocess+RF vs. SparseConvNet

Accuracy

Nb classes	RF	CNN
20	60.5	61.2
51	48.5	58.1

but low **quality** images and much **resizing**



SparseConvNet + Zooscan

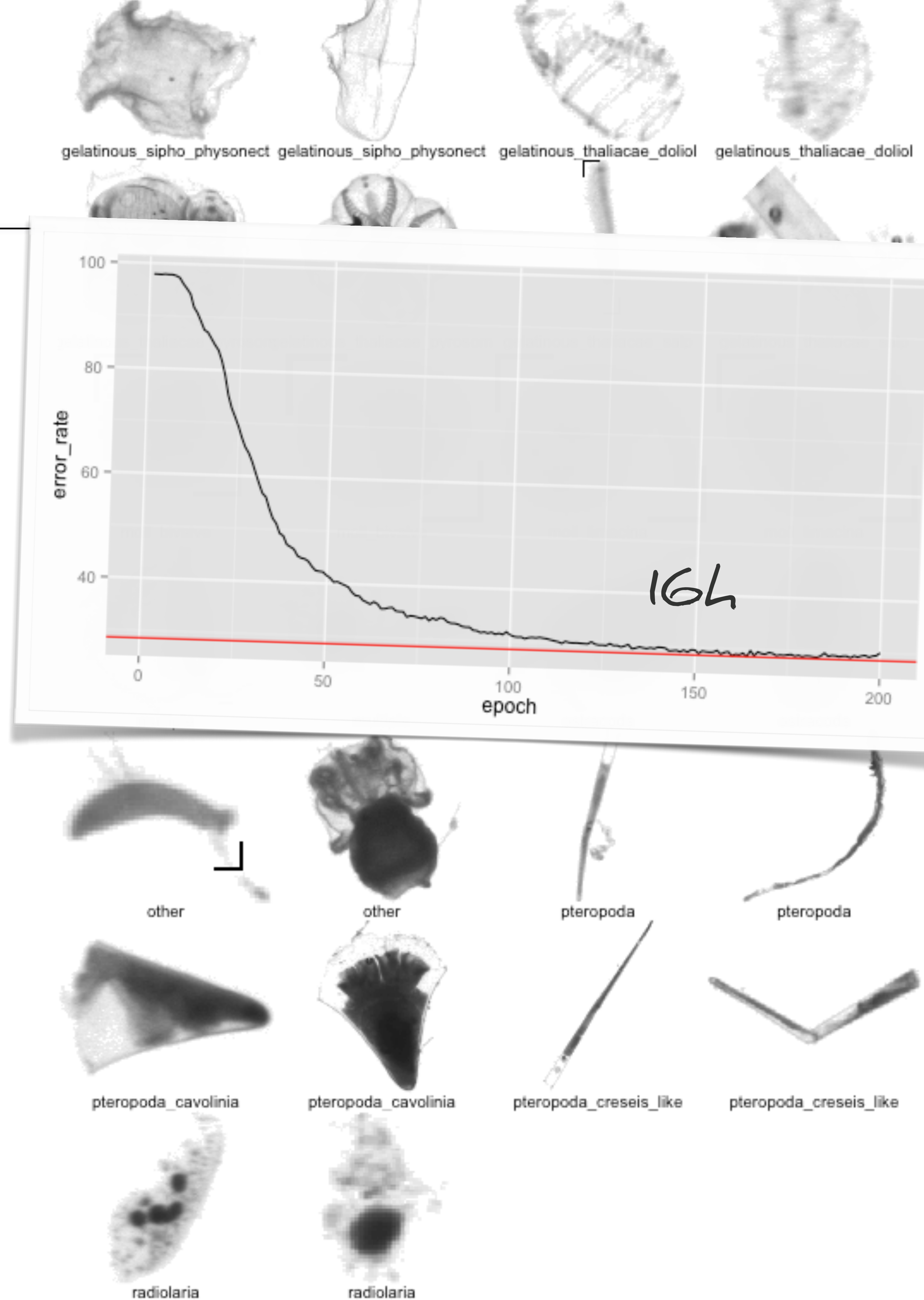
10k images for training, 80k for testing

Zooprocess+RF vs. SparseConvNet

Accuracy

Nb classes	RF	CNN
20	60.5	61.2
51	48.5	58.1

but low **quality** images and much **resizing**



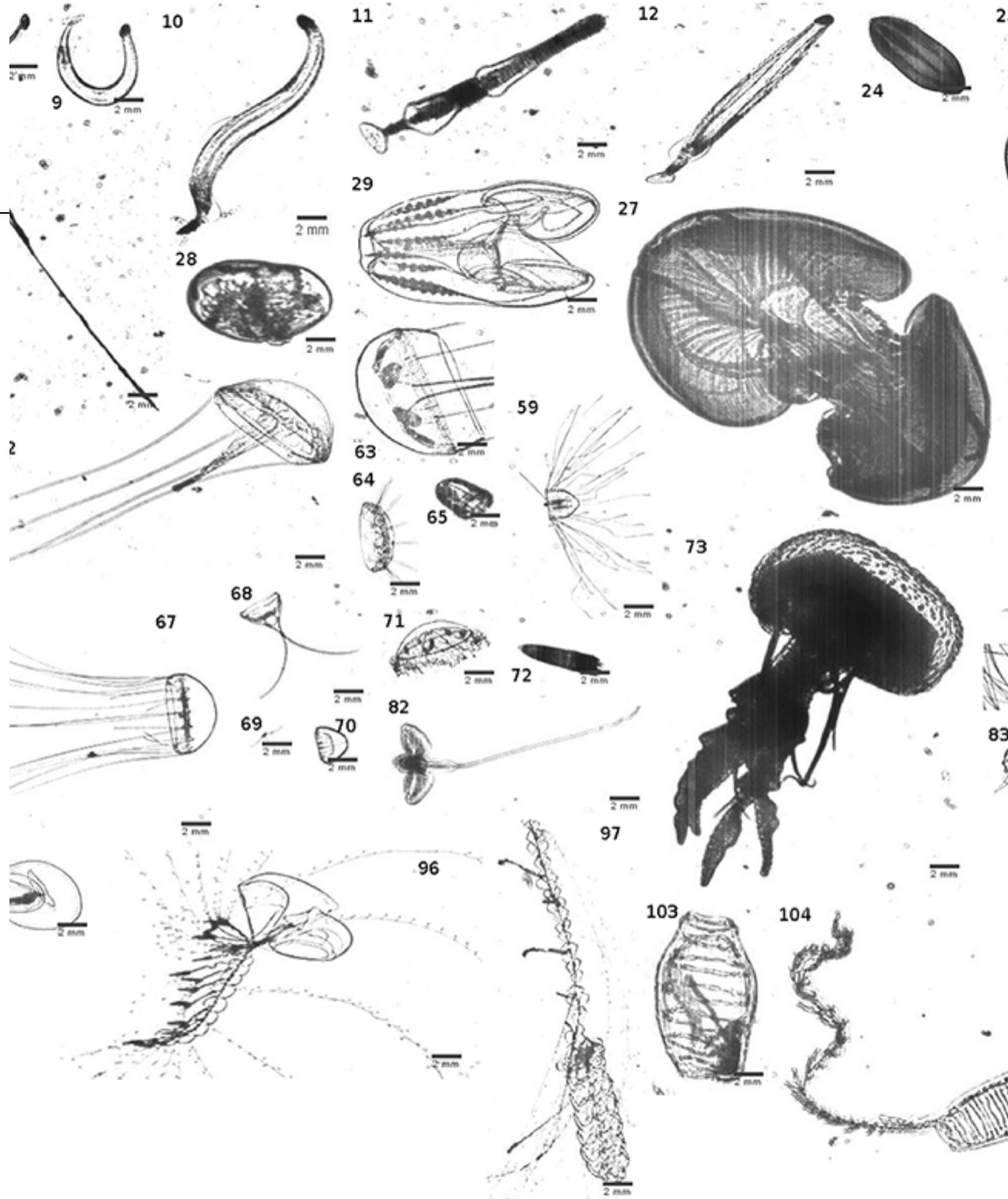
SparseConvNet + ISIIS

50k images training

24M images in **120** classes; 75k tested

87% overall accuracy

But the most biologically interesting classes are the **rare** ones (and accuracy is lower for those)



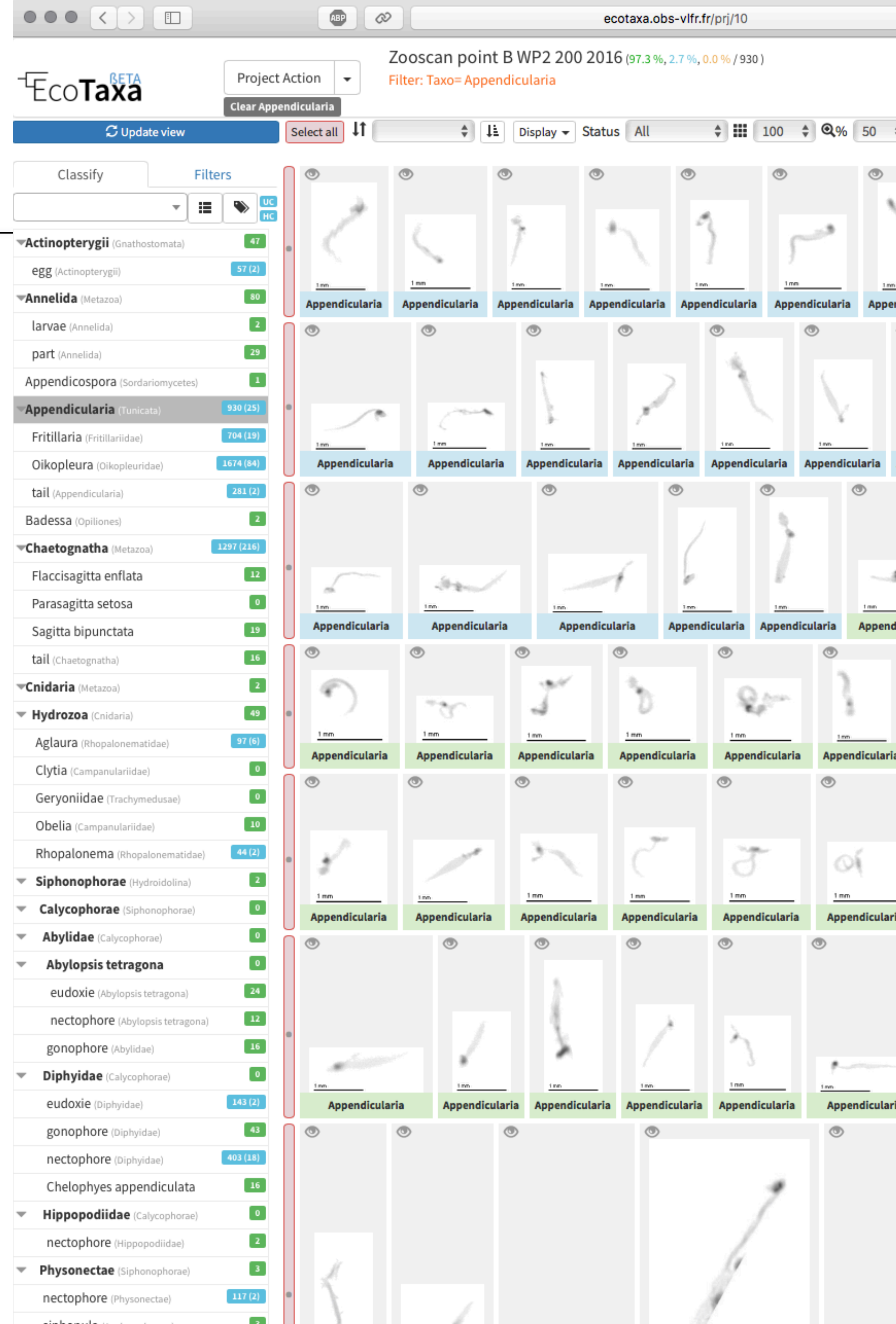
Current plans

CNNs **limitations**

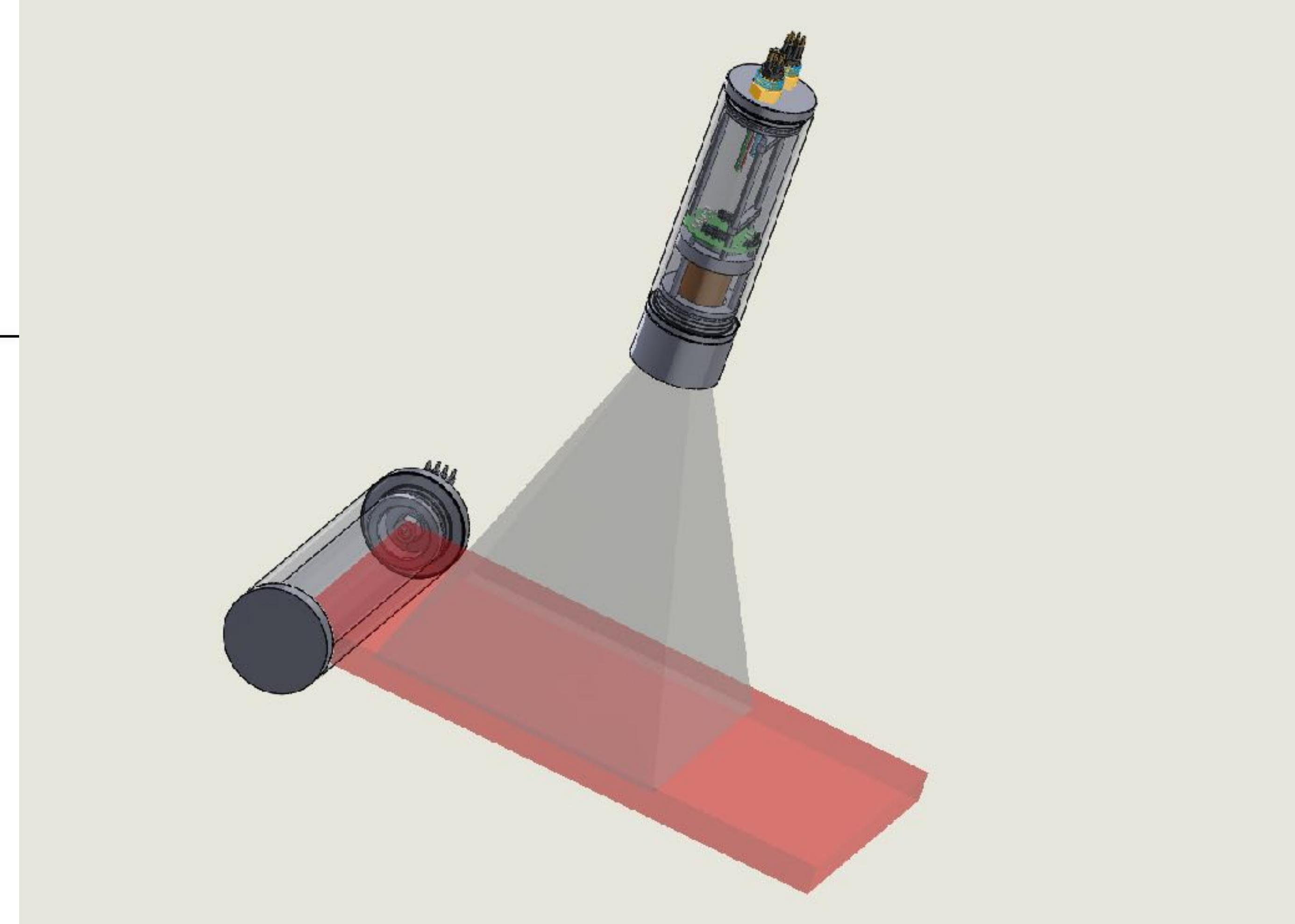
- only **image** data
- **slow**
- needs **lots** of training data

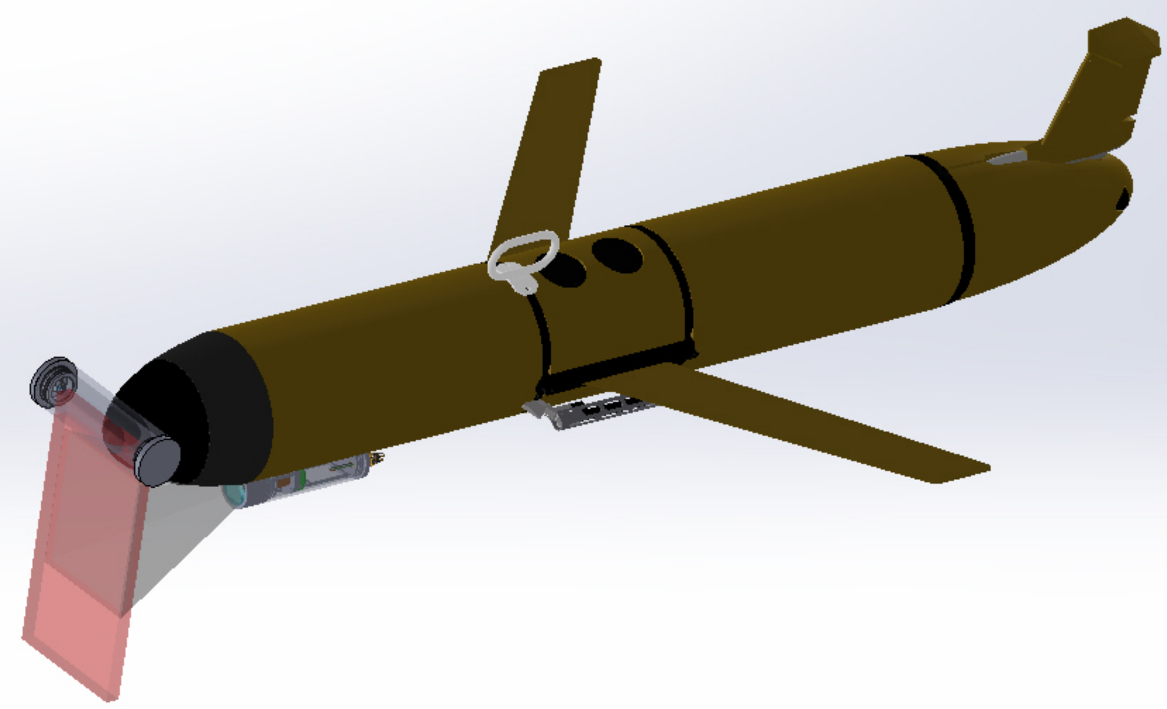
Extract **features** from CNN, combine them with **size** (and **metadata**), train RandomForrest

Integrate this as a "**one-click**" solution in a web application for plancton image classification



Future challenges



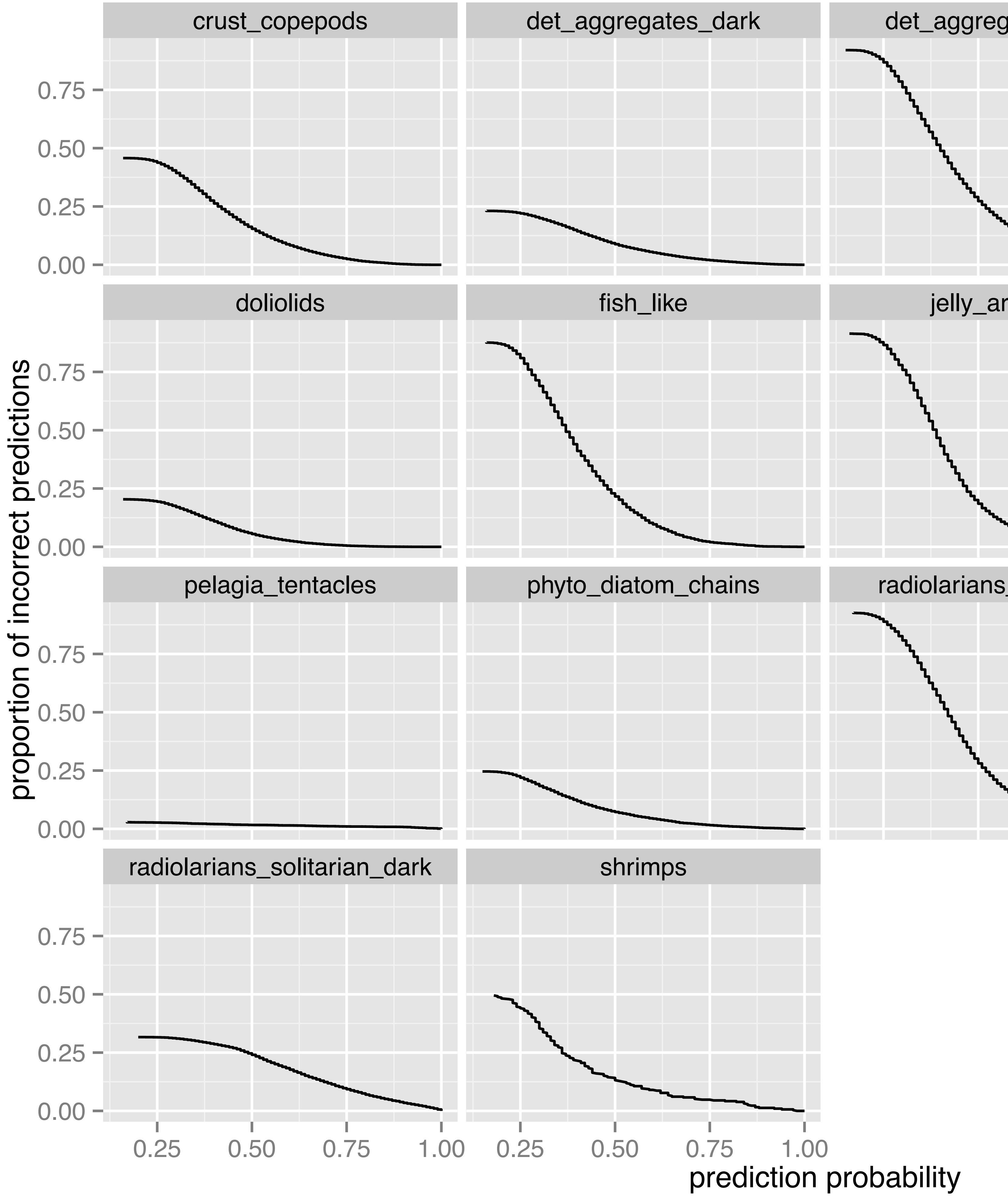


New **smart** sensors

- Take the **image** and extract "particles"
- Need to send data in real time
- Classification needs to be done **inline**, with **little** power (0.1W at 0.1fps)

One more thing...

Classification score



All algorithms produce a **score**, not a Yes/No answer

What if we could **throw out** bad scores?

Classification score

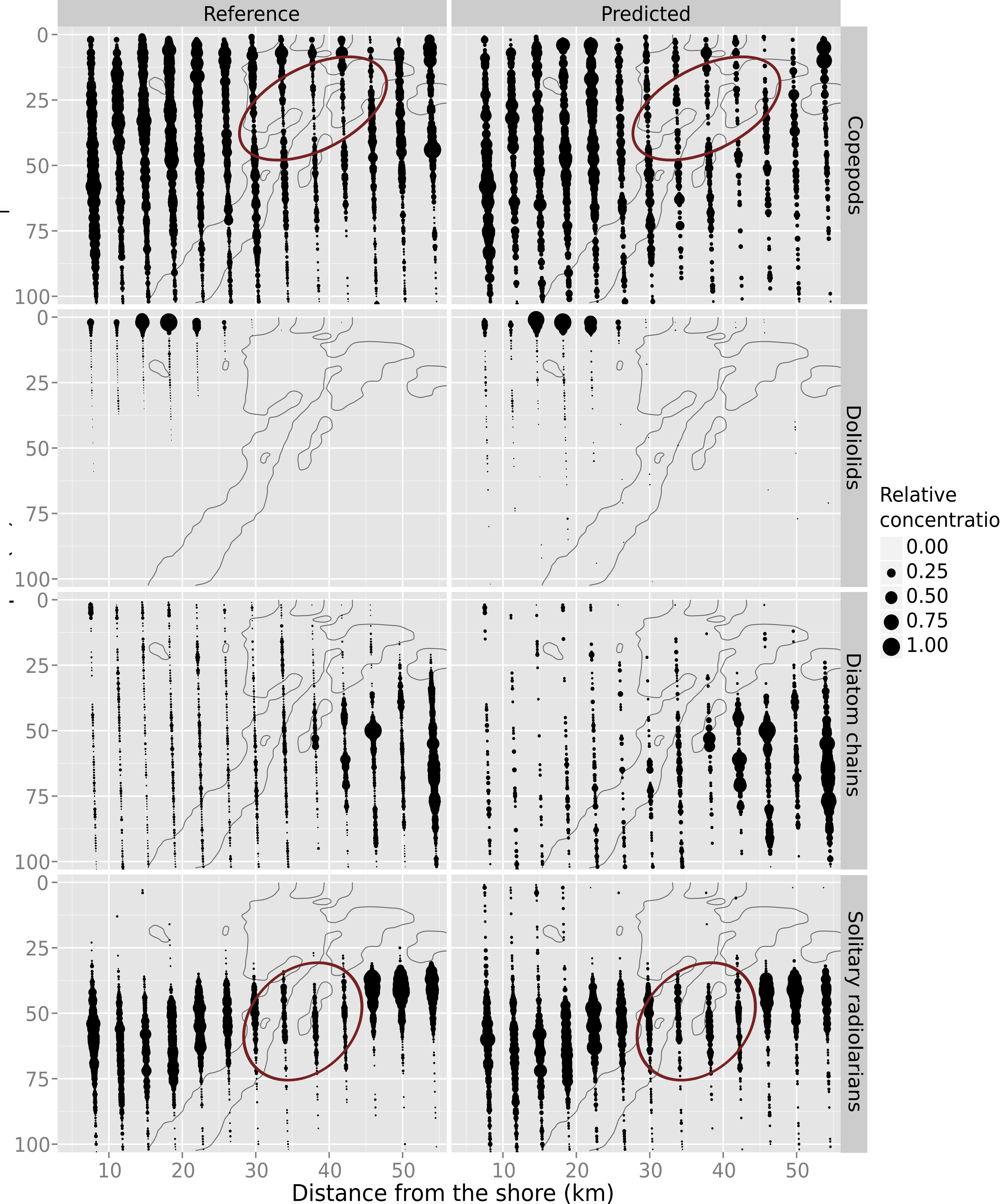
151k images of organisms

All automatically **and** manually sorted

Set score **thresholds** to reach 99% precision

⟹ discard 70% of objects

Compare the reference, full dataset and the thresholded one



Classification score

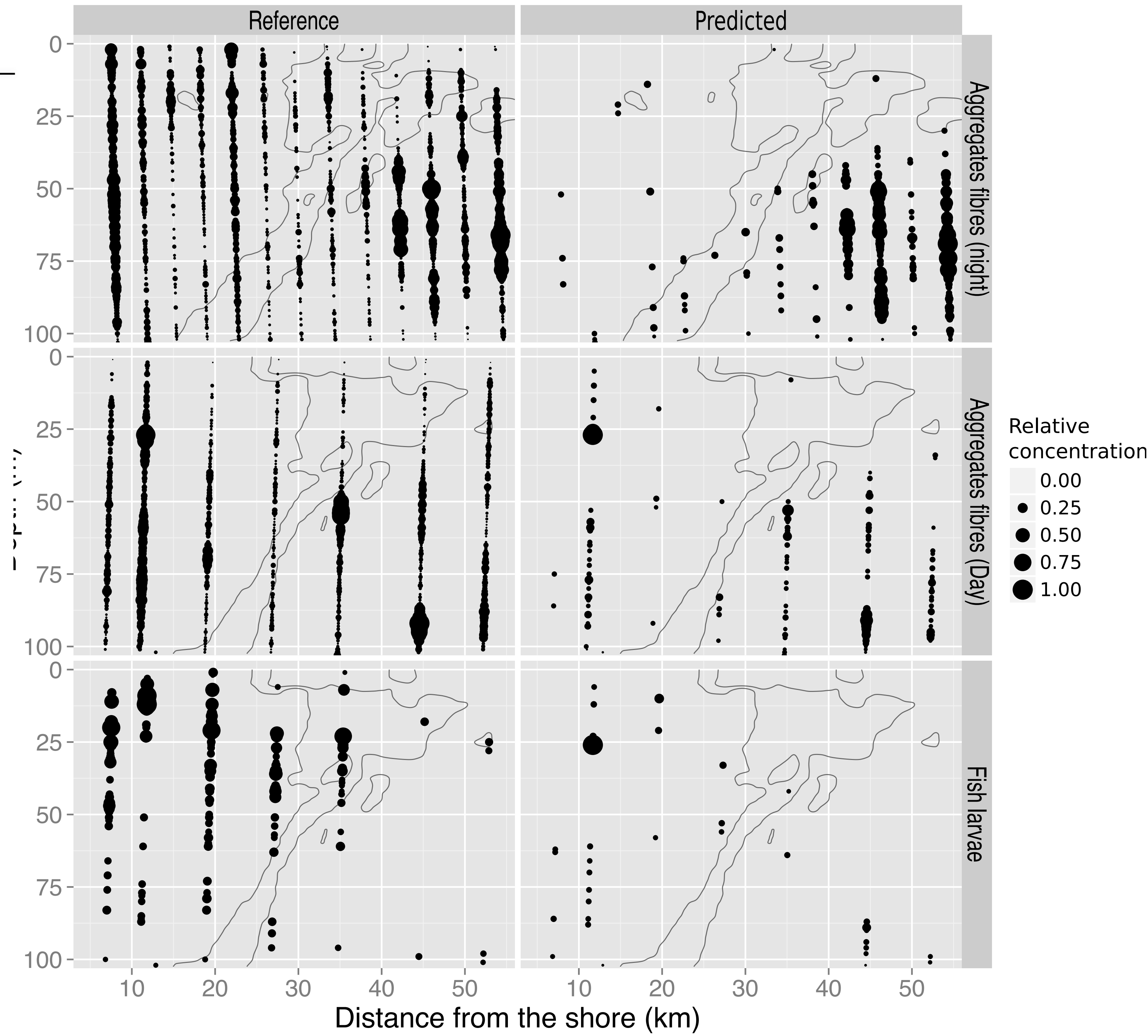
151k images of organisms

All automatically **and** manually sorted

Set score **thresholds** to reach 99% precision

⟹ discard 70% of objects

Compare the reference, full dataset and the thresholded one



Classification score

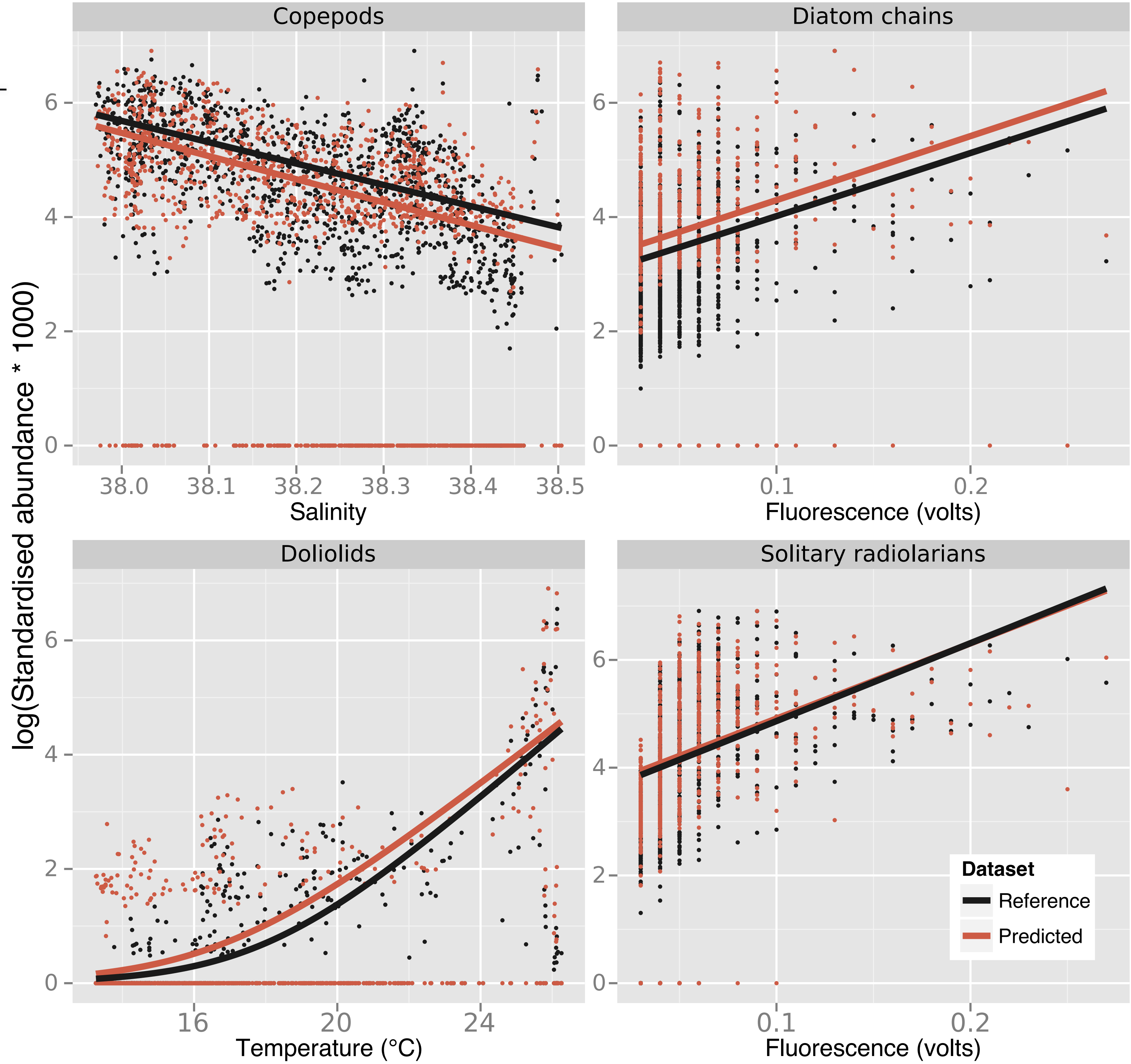
151k images of organisms

All automatically **and** manually sorted

Set score **thresholds** to reach 99% precision

$\Longrightarrow$ discard 70% of objects

Compare the reference, full dataset and the thresholded one




Merci pour
votre attention