From classification to cetaceans tracking by Passive Acoustic and AI Frameworks

Sébastien Paris[1,2,4] *Hervé Glotin*[1,2,4], *Lilou Dantin*[1,2,3,4], *Pascale Giraudet*[1,2,4], *Adeline Paiement*[1,2,4], *Stéphane Jespers*[1]

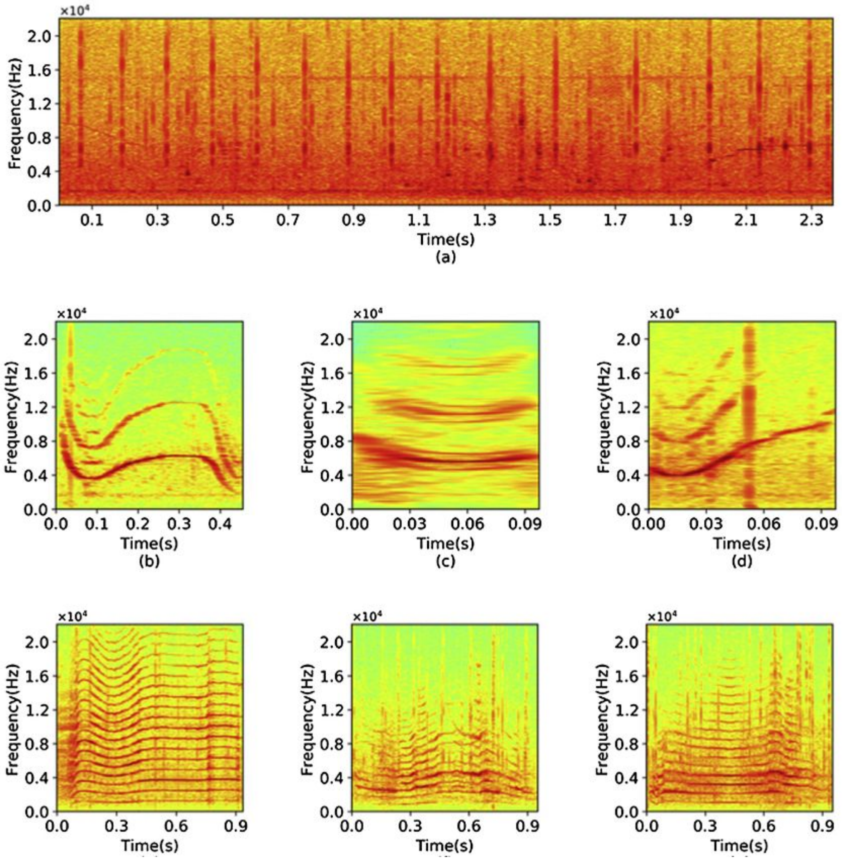
[1] *Centre International d'Intelligence Artificielle en Acoustique Naturelle*

[2] *Laboratoire d'Informatique et des Systèmes, CNRS, Université de Toulon*

[3] *Parc National de Port-Cros*

[4] *Chaire IA ADSIL AID DGA ANR-20-CHIA-0014*

What type of acoustic signals are emitted by marine mammals ?



Spectrogrammes (Représentations temps-fréquences)

Clics → Echolocation

Whistles

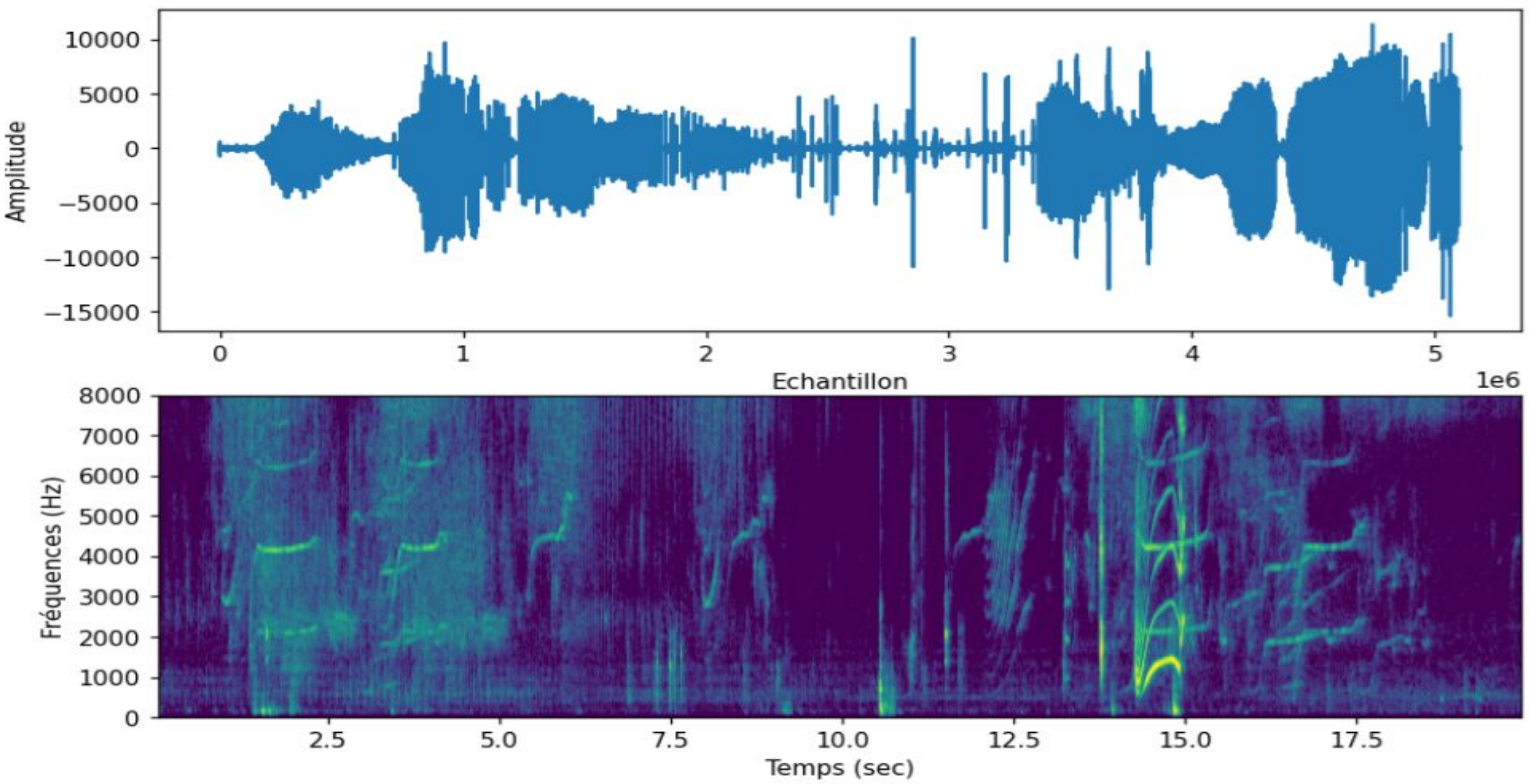
→ Socialization & Communication

Pulsed "calls"

Everything together:
Huge Cocktail party !!!

Jiang, J., et al,. (2019). Study of the relationship between pilot whale (Globicephala melas) behaviour and the ambiguity function of its sounds. *Applied Acoustics*, *146*, 31-37.

Everything together: huge Cocktail party !!!



Main motivations from our bioacoustic works (from 2000...)

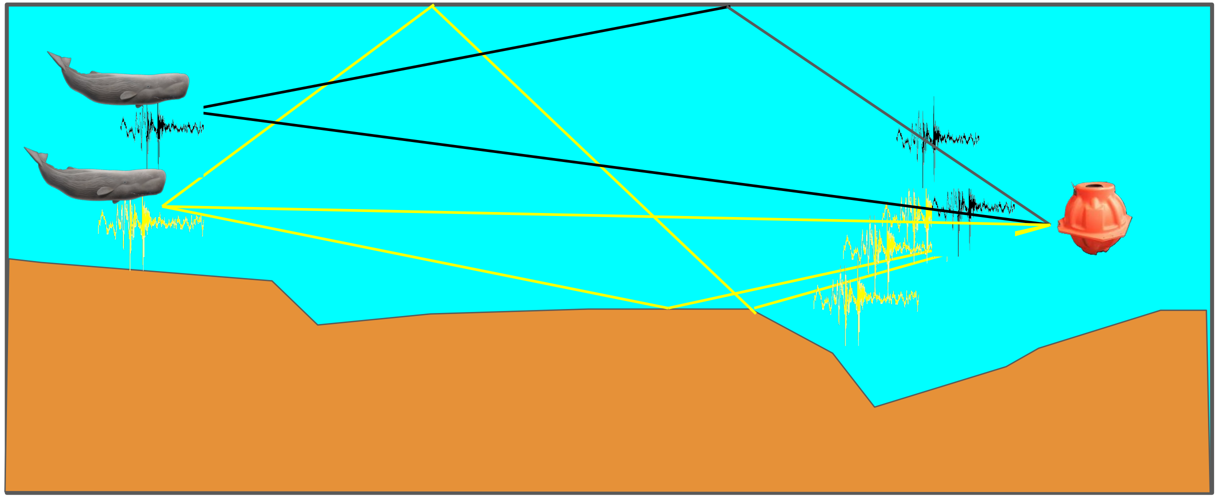
Given some collected underwater acoustic data **in a passive way** (mostly unsupervised), we are working (since decades) on these 5 different tasks:

1 - Detection : Is there at least one animal surrounding the sonobuoy ?
2 - Classification: What species have been detected ?
3 - Sequence modeling : What mammals are trying to say ? (communication understanding)
4 - Tracking : Where mammals are ?
5 - Optimal control/Reinforcement Learning: Where to deploy our sonobuoy ? (to maximize the last four tasks performances)

→ **Automatic tool to output biopopulation indicators**

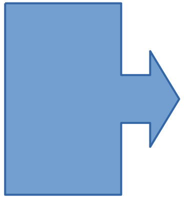
A common denominator for all these tasks: we went from signal processing/statistical modeling to some (full) machine learning (ML)/artificial intelligence (AI) solutions....

Underwater acoustic channel



$$r_i(t) = (g(s(t) * h_i(t) + n(t)) * a_i(t) + b(t)$$

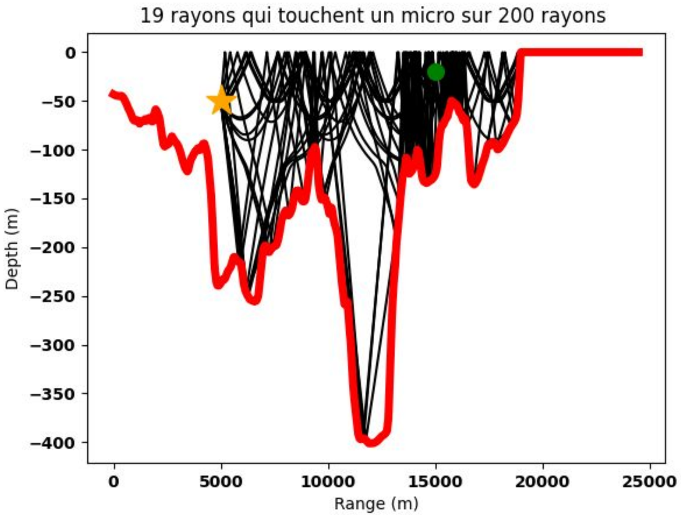
s(t) : source signal (calls, clicks, ...)
$h_i(t)$: propagation/scattering equivalent transfer function
g(t) : transmission loss
n(t) : ambient sea noise
$a_i(t)$: hydrophone transfer function
b(t) : receiver noise
$r_i(t)$: observed signal on hydrophone i



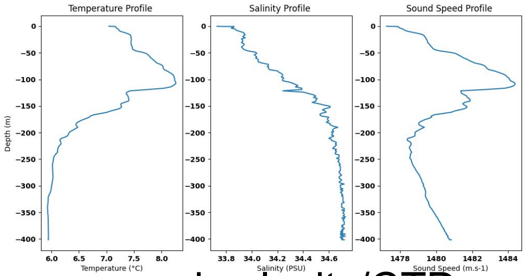
Very complex and noisy signal

Just to give an idea of the channel complexity

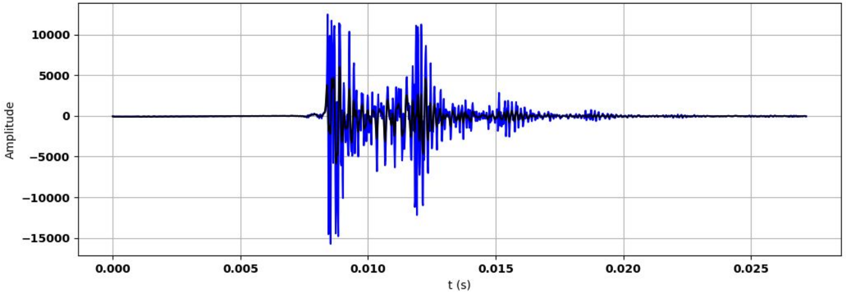
The sound propagation involves many physical aspects : *reflexion, refraction, diffraction, back-scattering, etc*… and depends a lot of parameters: *frequency, bathymetry, pressure, temperature, soil regularity, etc….*



ray-tracing engine



sound velocity/CTD

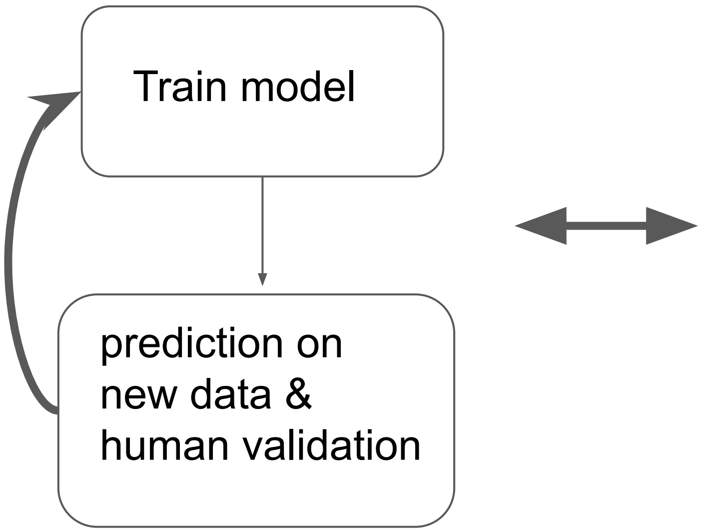


emitted *s(t)* (blue) and transmitted (black) acoustic signals *h(t)*s(t)*

Using AI in bioacoustic : what was (and still is) the more challenging ?

GET LABELS/GROUND TRUTH !!!!! (especially for task 4 in PAM framework)

- We started with just hundreds of examples in total: highly unbalanced and with a lot of label noise



- Starting with mostly unsupervised techniques
- took years to have acceptable results

From signal processing to statistical learning (< 2013)

- At least for **tasks 1-2**, from 2006-2007 => more datasets available (with partial labeling),
- we started to work on (mostly) unsupervised ML technics to produce **latent representations**

$$\mathbf{z} = f_\theta(l(\mathbf{r};\beta))$$

where *l* can be typically a TF representation (*STFT, MELcep, scalogram*, etc..) with fixed β hyper-parameter and θ is the trained non-supervised representation. Among them:

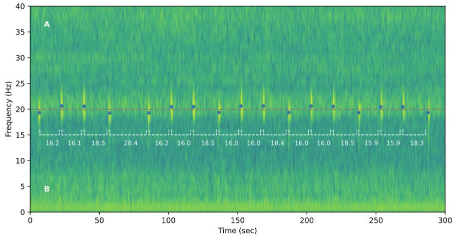
- *clustering/Bag Of ..*
- *GMM*
- *sparse coding+dictionary learning*
- *Fisher vectors*
- *etc…*
- Can be considered of a first trained hybrid learned representation
- Improved a lot performances for tasks 1-2

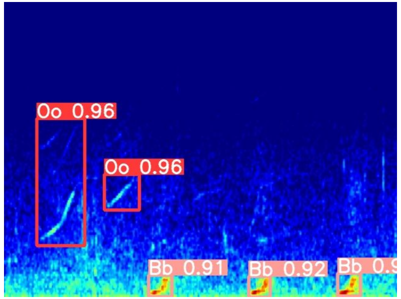
From 2013 for tasks 1-2

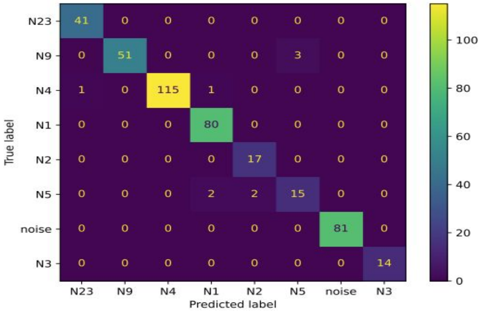
The IA's tsunami began. Better **latent representations** are obtained with modern *NN* architectures (*CNN, RNN, Unet, Transfomer, etc…*). Key points were:

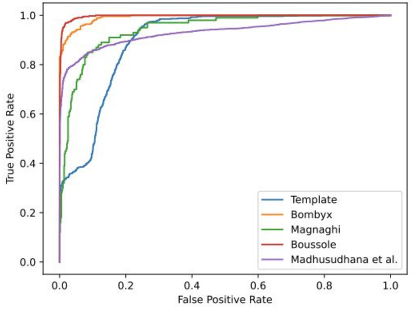
- Huge effort in labeling (partially) databases
- Better optimization gradient based solutions (*Adam, autodiff, etc...*),
- Transfer learning, self-supervised learning, active learning technics
- Regularization by data augmentation (noise, transform, *etc..*), dedicated layers

Whales detection/classification[1,2] with low-power CNN based architectures



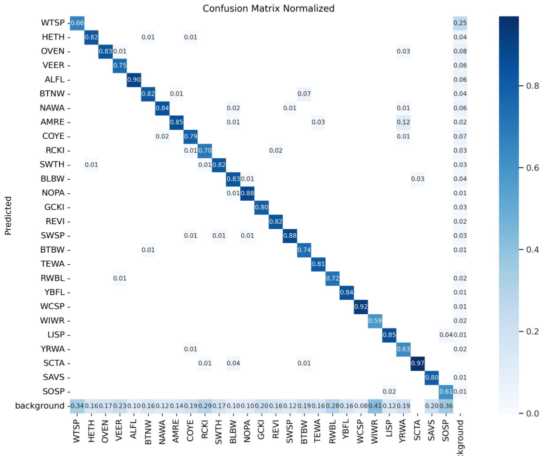




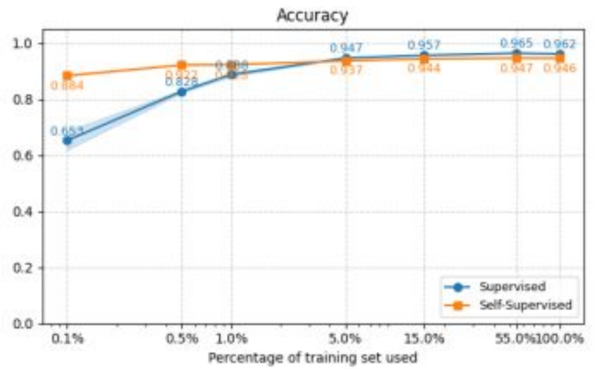


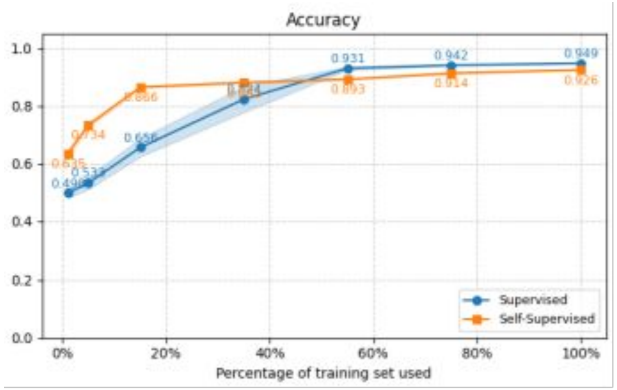
[1] Paul Best, *Automated Detection and Classification of Cetacean Acoustic Signals*, PhD Thesis, 2022
[2] Paul Best and al, *Temporal evolution of the Mediterranean fin whale song,* Scientific Report, 2022

- Birds classification[3] (TFR + preprocessing + YoLo V12)



- Fin whale detection[4] (SSL vs Supervised Transformer model)





[3] Stéphane Chavin, PhD Thesis, 2023-...
[4] Adam Chareyre and al, *Self-Supervised vs Supervised Representation Learning for Fin Whale Vocalization Detection*, Neurips, 2025

For tasks 1-2, job is (almost) done !

Take home message:

- Performances for tasks 1-2 are **now quiet good** (> 85% Acc for most datasets)
- More and more sequences are **automatically extracted**, **analysed and labelled** (> [10K-300K] detections per inference session)
- **In practice,** for tasks 1-2, fine-tuned YOLO Vx.. reaches ~SOTA even in cocktail party
- **In most of the case,** no really need cumbersome ultra advanced IA arsenal (low-energy embedded system incompatible)

Why AI also for tasks 4-5 ?

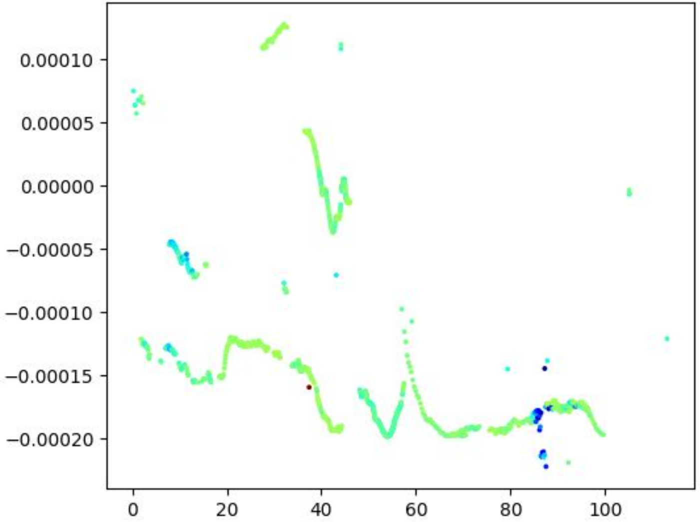
For **task 4**, with sonobuoy/hardware developpements we increased the :

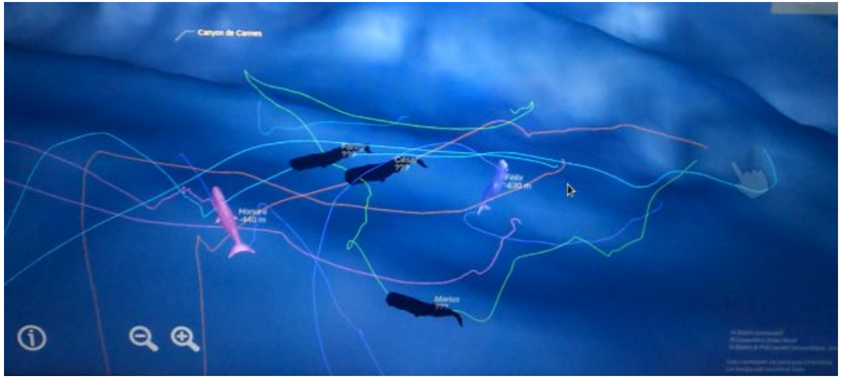
- number of hydrophones (up to 5)
- frequency sampling (up to 512 kHz)
- sensitivity/SNR

more robust/accurate TDOA estimators BUT CRLB shows **poor range estimators** from TDOA/TOA measurements.

1- direct localization approach : from TDOA's → $\hat{\mathbf{x}}_k = f^{-1}(\hat{\tau}_k)$

- hyperboloids intersection
- Weighted LLS



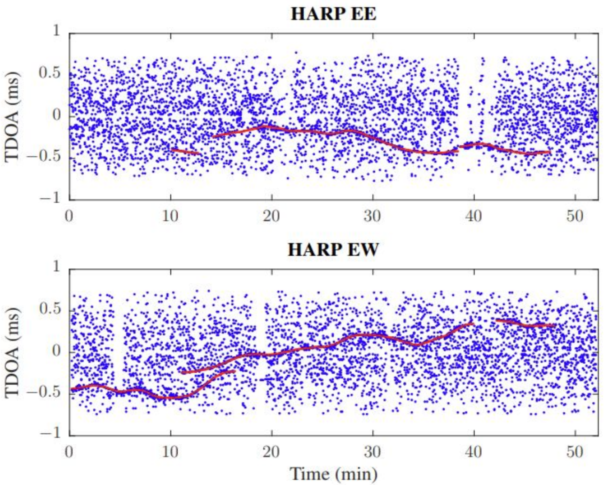


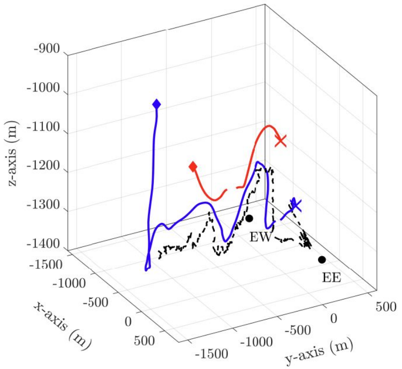
- need to remove clutter/ghosts TDOAs and
- isolate individual track.
- Can be done offline by unsupervised learning (advanced clustering GNN). Not yet fully automatic

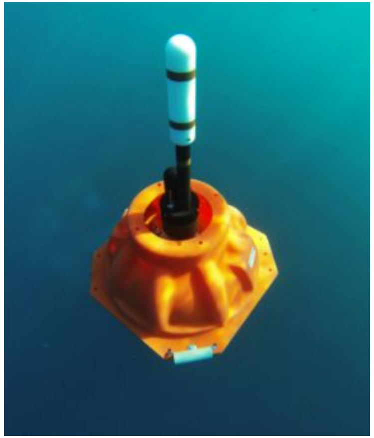
Sequential nonlinear filtering for MultiTarget Tracking

2- sequential tracking approach : given a sequence of *TDOA* *(or doppler,angle,range, etc)*.

from localization $\widehat{\mathbf{x}}_k^l = f^{-1}(\widehat{\tau}_k^l)$ → $\widehat{p}(\mathbf{x}_k^l | \widehat{\tau}_1, \ldots, \widehat{\tau}_k)$ to tracking





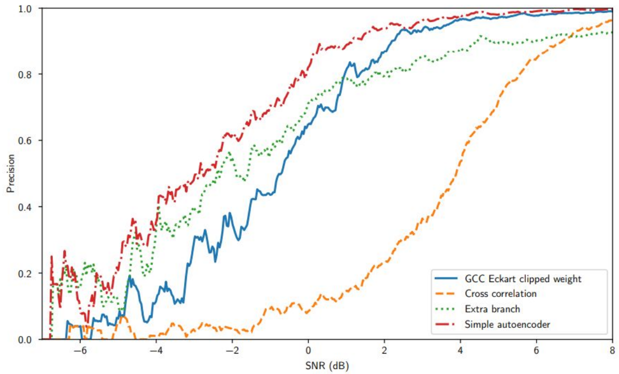


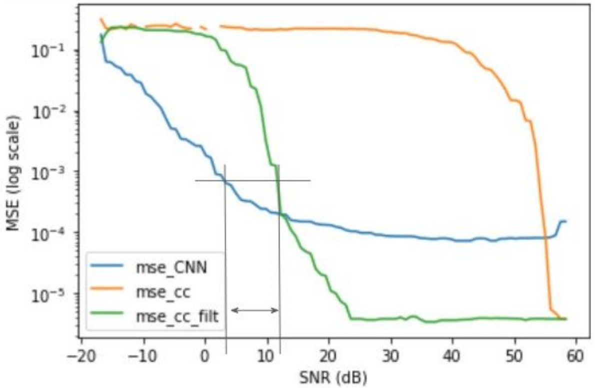
Main difficulty in MTT is the (combinatorial) **assignment problem between measures and targets** => *(P)MHT, JPDAF, Bayesian filter*[5]*, ect..*

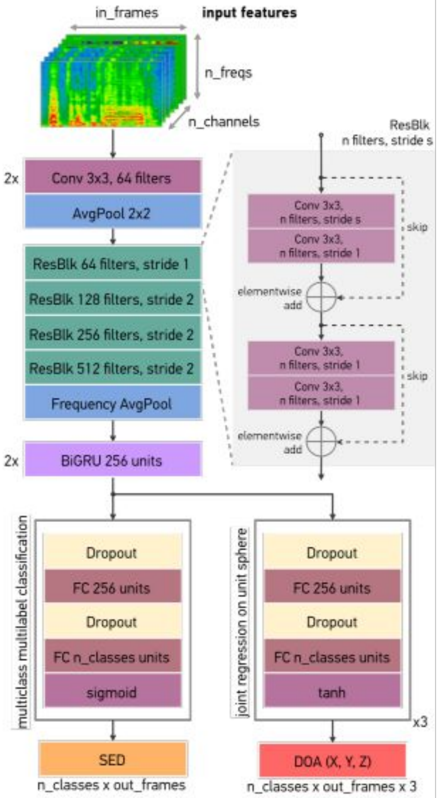
[5] J Jang, Bayesian Detection and Tracking of Odontocetes in 3-D from Their Echolocation Clicks, arXiv preprint arXiv:2210.12318

Coupling AI and MultiTarget Tracking

One way to overcome combinatory : train model robust *TDOA/DOA/range/angle* estimators[6] (even direct positioning) from sound events **with builtin source separation**[7]







Independent parallel filtering

$$\widehat{p}(\mathbf{x}_k^l|\widehat{\tau}_1^l,\ldots,\widehat{\tau}_k^l)$$

$$\widehat{p}(\mathbf{x}_k^l|\widehat{\mathbf{DOA}}_1^l,\ldots,\widehat{\mathbf{DOA}}_k^l) \quad \widehat{\mathbf{DOA}}_k^l = f_{\hat{\theta}}(\mathbf{r_k})$$

[6] Maxence Ferrari, *Study of a Biosonar Based on the Modeling …..*, PhD Thesis, 2020
[7] T. Nguyen, *Spatial Cue-Augmented Log-Spectrogram Features for Polyphonic Sound Event Localization and Detection, IEEE Trans ASLP*

New framework : Multi-Target Tracking with Transformer

3 - With Transformer like we can train **directly** (acoustic) sequences to (trajectories) sequences

$$\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n), \mathbf{z}_i \in \mathbb{R}^p$$

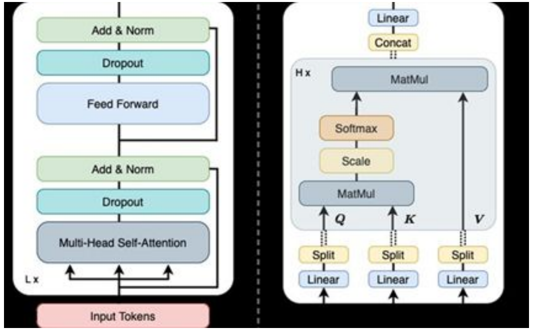
(eg. embedding from signals per hydrophone)

Transformer $\mathbf{X} = g_{\hat{\theta}}(\mathbf{Z})$

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n), \mathbf{x}_i \in \mathbb{R}^v$$

(eg. animal's position)

Attention layer $\mathbf{x}_i = \mathbf{W}_O \left(\sum_{j=1}^{n} \alpha_{i,j} \mathbf{W}_V \mathbf{z}_j \right)$ $\alpha_{i,j} = \dfrac{\exp\left(\frac{(\mathbf{W}_Q \mathbf{z}_i)(\mathbf{W}_K \mathbf{z}_j)}{\sqrt{d_k}}\right)}{\sum_{k=0}^{n} \exp\left(\frac{(\mathbf{W}_Q \mathbf{z}_i)(\mathbf{W}_K \mathbf{z}_k)}{\sqrt{d_k}}\right)}$

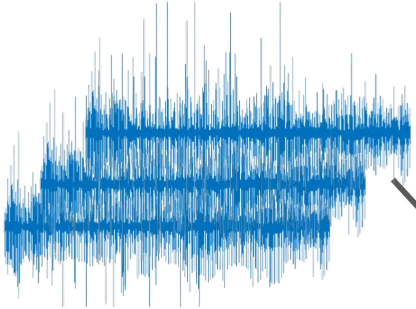


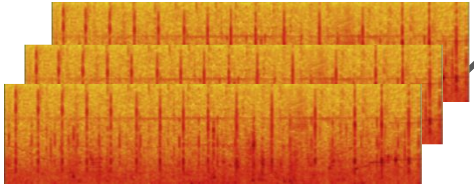
Passive Acoustic Tracking with Transformer

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_K\} = g_{\hat{\theta}}(\{\mathbf{z}_1, \ldots, \mathbf{z}_K\})$$

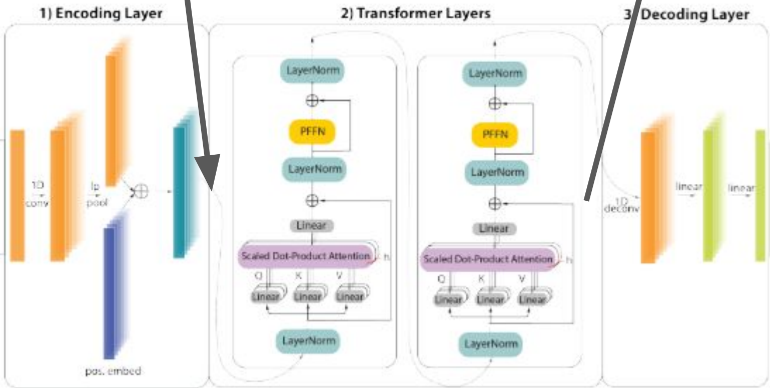
$$\{\mathbf{z}_1, \ldots, \mathbf{z}_K\} \quad \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$$

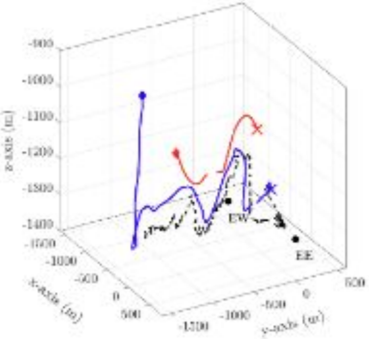
$$\mathcal{L}(\hat{\mathbf{r}}_k, \mathbf{r}_k)$$





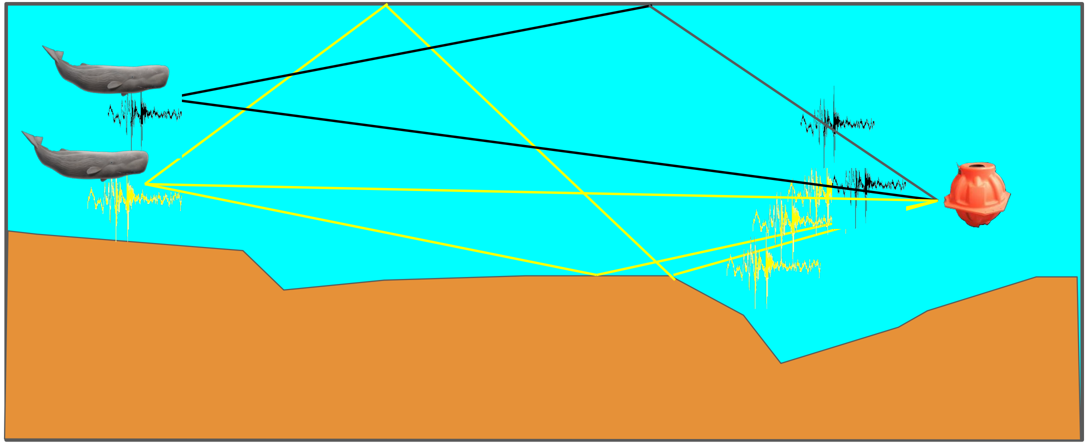
ViT

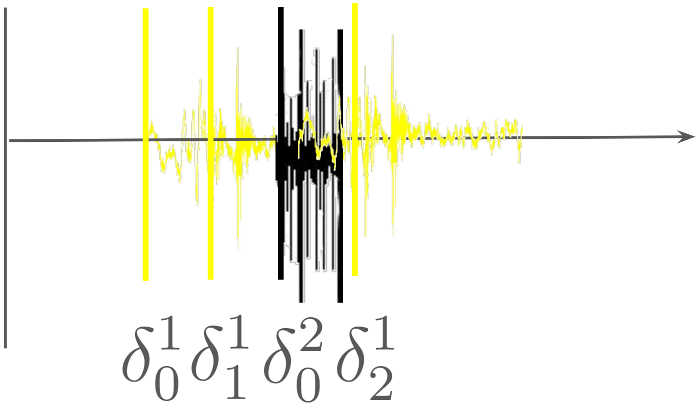




What the representation must learn via Transformer ?

Answer: the underlying source separation problem (animals, echoes, etc..)

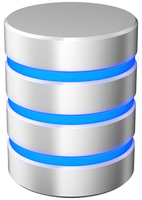
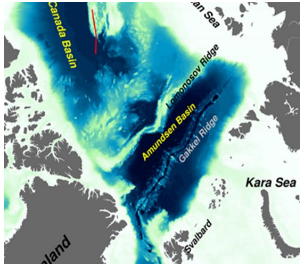


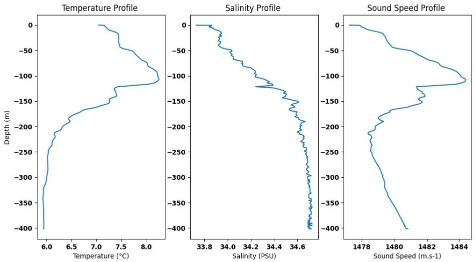


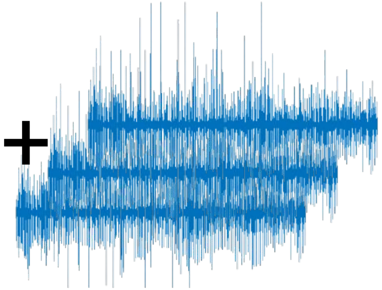
We need a dataset dedicated to PAT !!!!

- Whatever tracking with 1/2/3 approach**, we need ground truth data** with **acoustic data** (A) and animal’s **trajectories** (T) to train models.

- Few datasets are available with all these informations together.

 =  + 

+ 

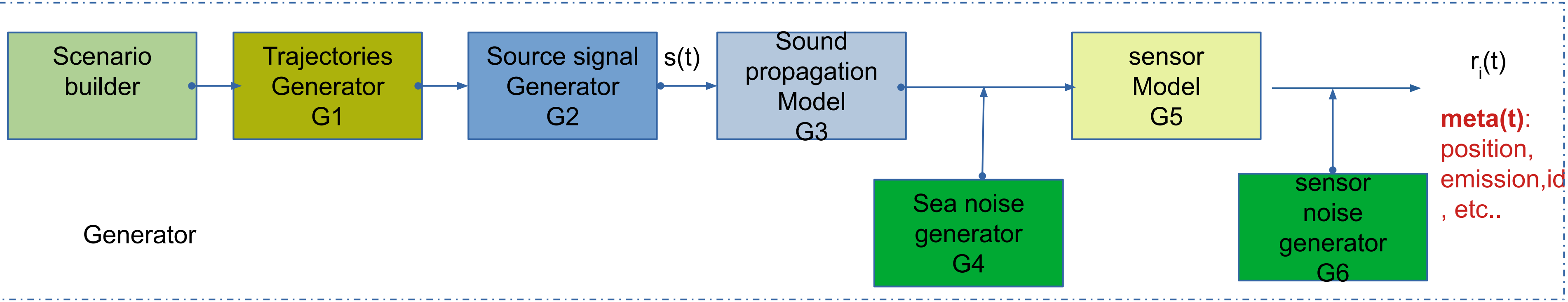
+ 

- **We need a digital twin/serious game of marine mammals** to generate realistic data

SeGaMas - Generator -

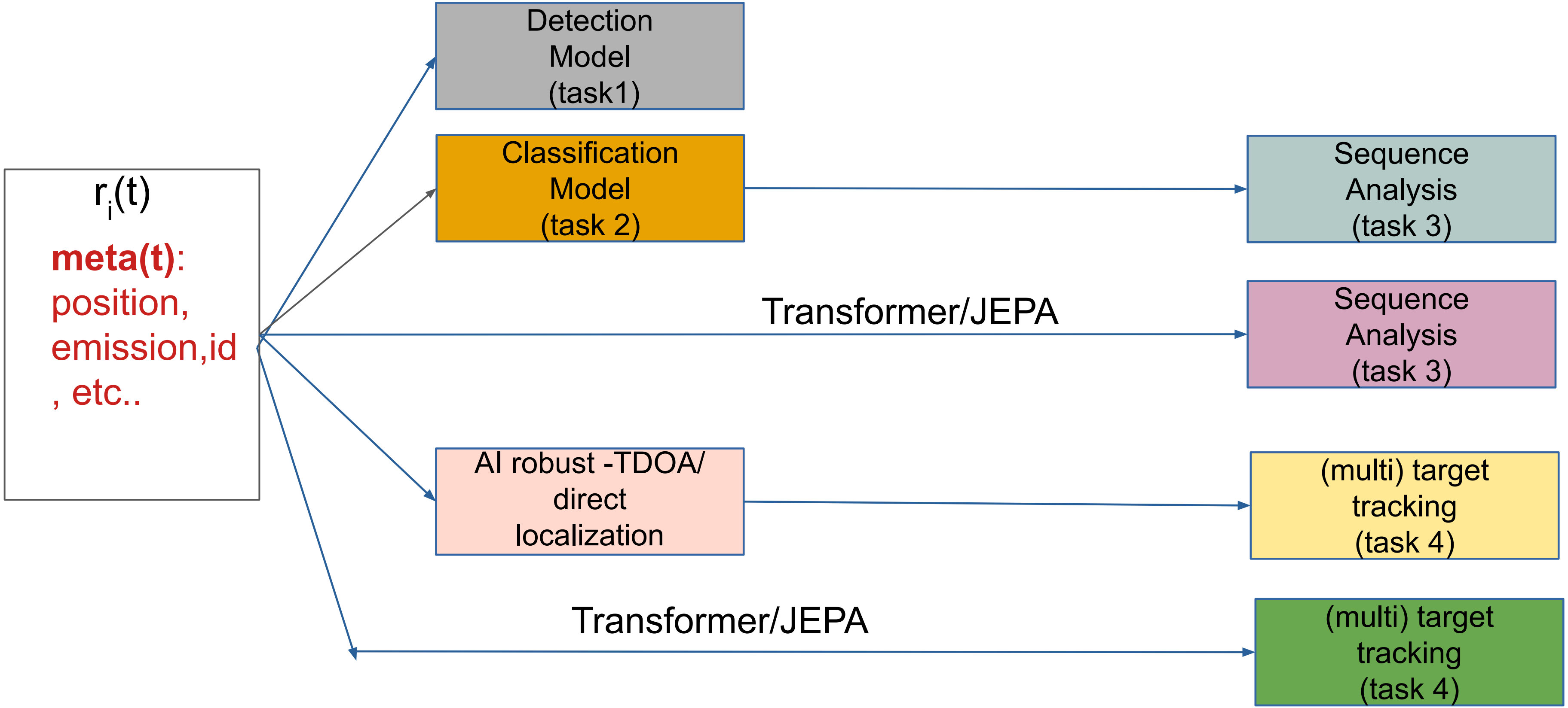
We started to build a complete **serious game** (L. Dantin 2025-) divided in two parts **: generator & trainers**. The generator has to:

- generate realistic mammals trajectories (cinematic, behavior, ROI, weather, food, multiple animals, etc...)
- generate realistic source emissions
- model sound propagation and sea noise characteristics
- model sonobuoy geometry and sensor characteristics



With SeGaMas generator, the goal is not only to generate realistic acoustic signals but **also all important associated meta-data/labels** for tasks 2-3-4-5

SeGaMas - Training models

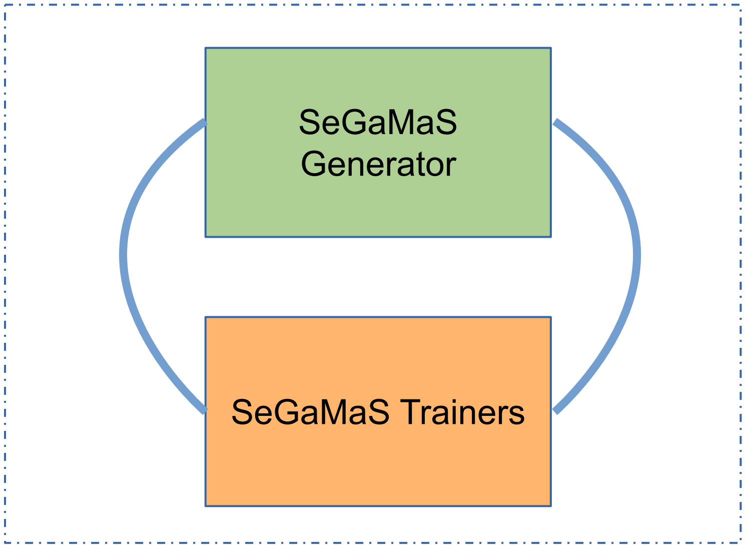


SeGaMas - Generator + Trainers

For **task 5** , thanks to all generated trajectories and associated sound events & meta labels, we can imagine find the best sensor's location minimizing such loss

$$L(\boldsymbol{U}) = \min_{\boldsymbol{U}} \{ E_{\boldsymbol{T}} [\sum_{k} det (cov (\boldsymbol{x}_k | \boldsymbol{Z}_{1:k}(\boldsymbol{U})))] \}$$

- $L(\boldsymbol{U})$ ← Sensor's location
- $E_{\boldsymbol{T}}$ ← Trajectories from G1
- $det(cov(\cdot))$ ← MTT (task 4) or PCRB



L(**U**) can be optimized by stochastic optimization technics or via RL (agent = sonobuoy)
Would be interesting to compare both way to solve the corresponding problem

Thanks for your attention !!
Questions ?