gipsa-lab

DGA

LabEx OSUG@2020

Observatoire des Sciences de l'Univers de Grenoble

Ecole doctorale
Terre Univers
Environnement

Data Institute
Univ. Grenoble Alpes

financé par
IDEX Université
Grenoble Alpes

Automatic classification of natural signals for environmental monitoring

PhD Defense, by Marielle Malfante

- Supervision - Jérôme Mars, Mauro Dalla Mura

- Funding - 50% DGA - 50% OSUG@2020

Volcanoes Monitoring

WHY?

- 1500 active volcanoes, 29.000.000 people < 10km of volcanoes, 50 to 60 eruptions / year,
- Societal / Economical impact,
- Scientific research.

HOW?

- Deformation, visual indicators, chemical composition, **seismicity**, etc,
- **Multimodal** problem,
- **Big Data.**

Flow of seismic data to be analyzed $f_s = 50Hz$

Seas & Oceans Monitoring

WHY?

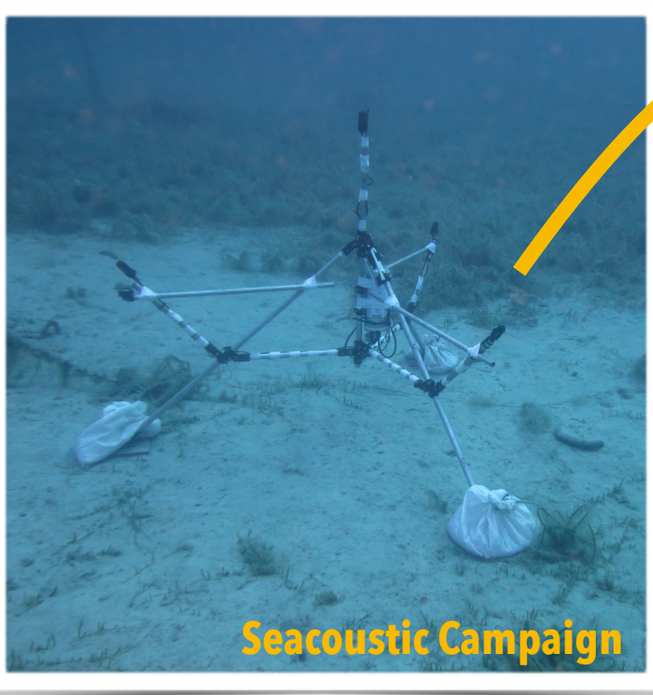
- 70% of Earth covered by seas, >90% of Earth biomass underwater,
- Scientific research (Earth`**big unknown'**),
- Focus on coastal areas (depth<300m) : resources (energy, fishery), submarines, etc.

HOW?

- **Data gathering campaigns**: drifting gliders, boats, specific sensors, etc,
- Multimodal problem: pressure, temperature, salinity, **acoustic** (**passive** or active) etc.

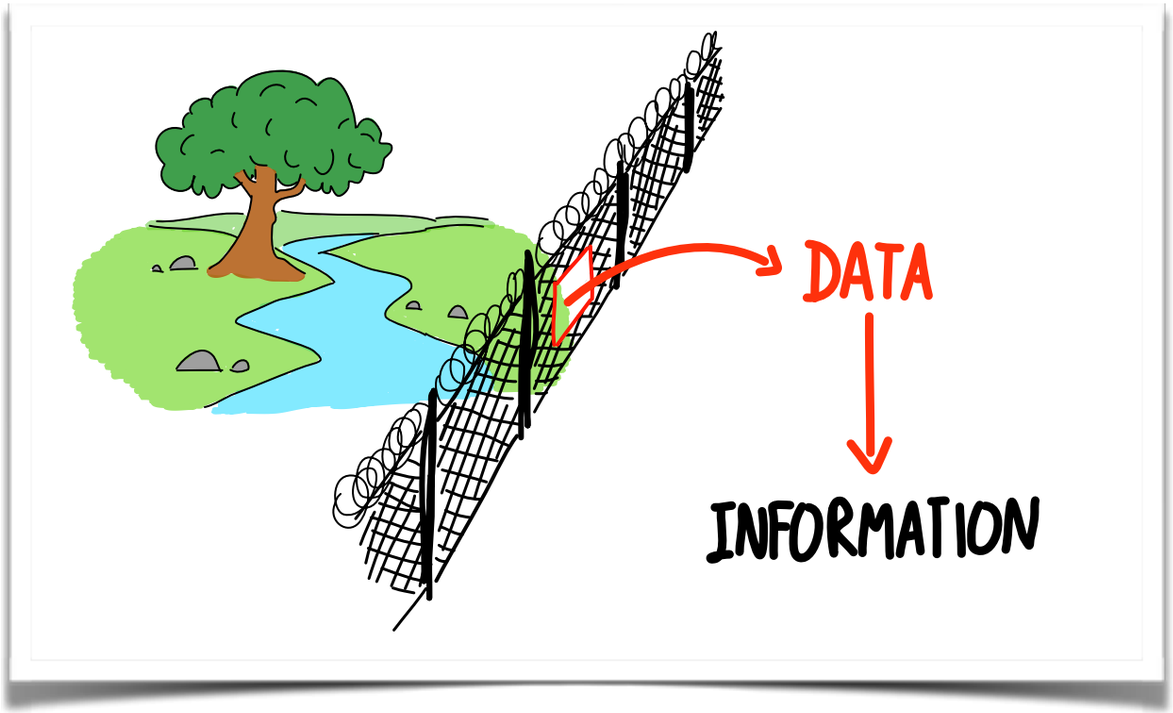

Flip boat


Fastwave seaglider


Seacoustic Campaign

Days of acoustic signals to be analyzed
$f_s = 156kHz$

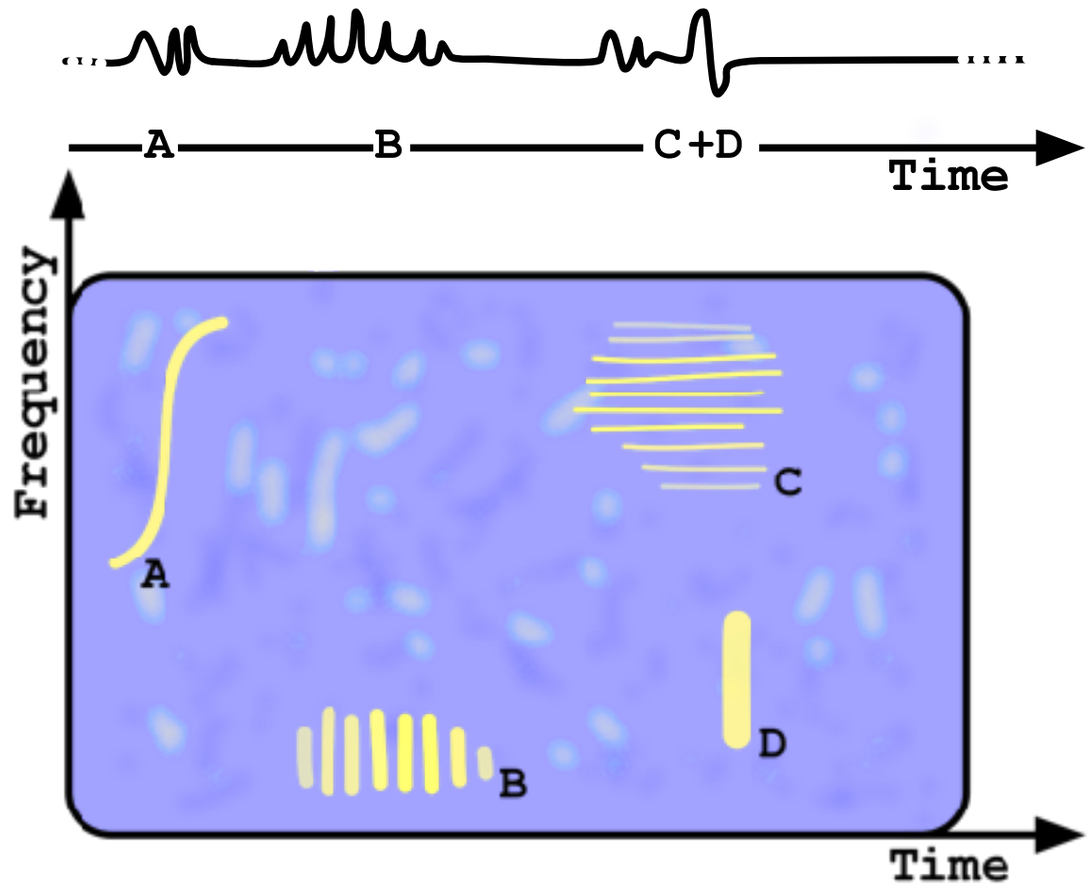
2 topics, 1 modelization



Environment	Volcanoes	Underwater areas
Wave	Seismic	Acoustic
f_s	50Hz	156kHz
Time scale	From seconds to weeks	From milliseconds to a few seconds

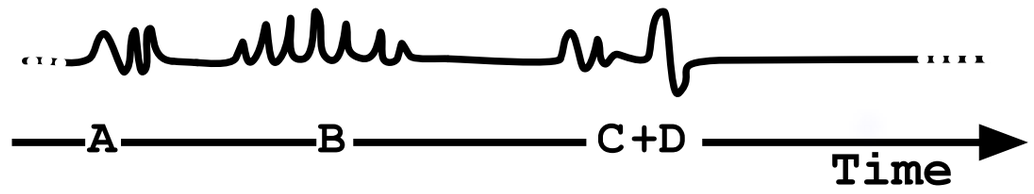
DATA?

- **Recording** = `**Continuous**' signal $[w_k]_{k=0}^{n_t}$

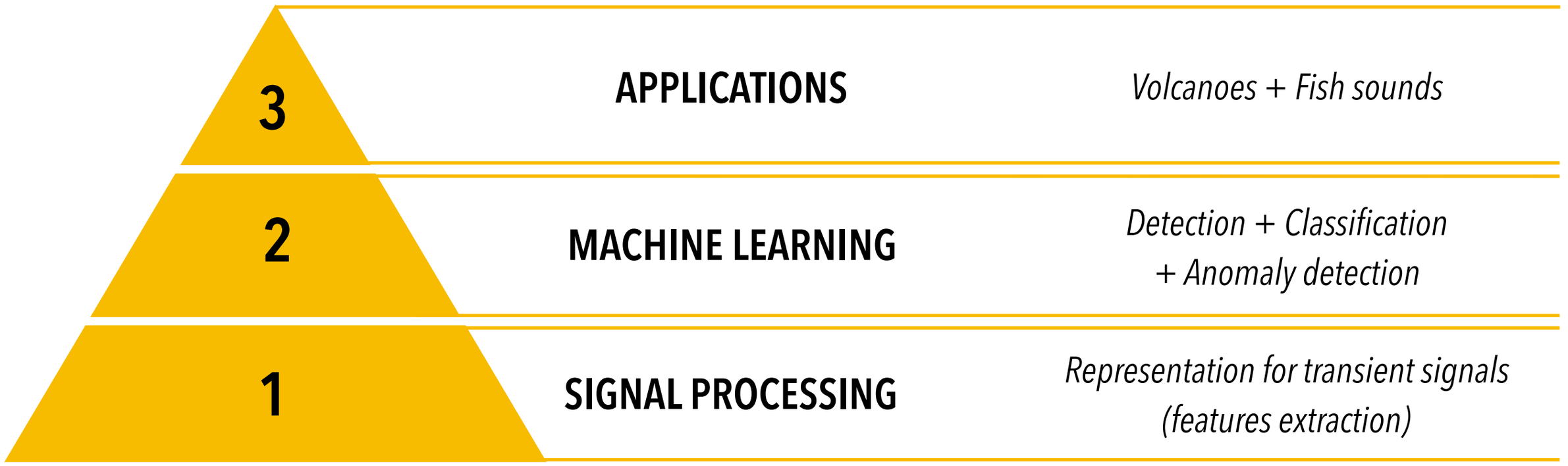


- **Amount of data** ⟶ **Automatic methods**

Issue & Contributions



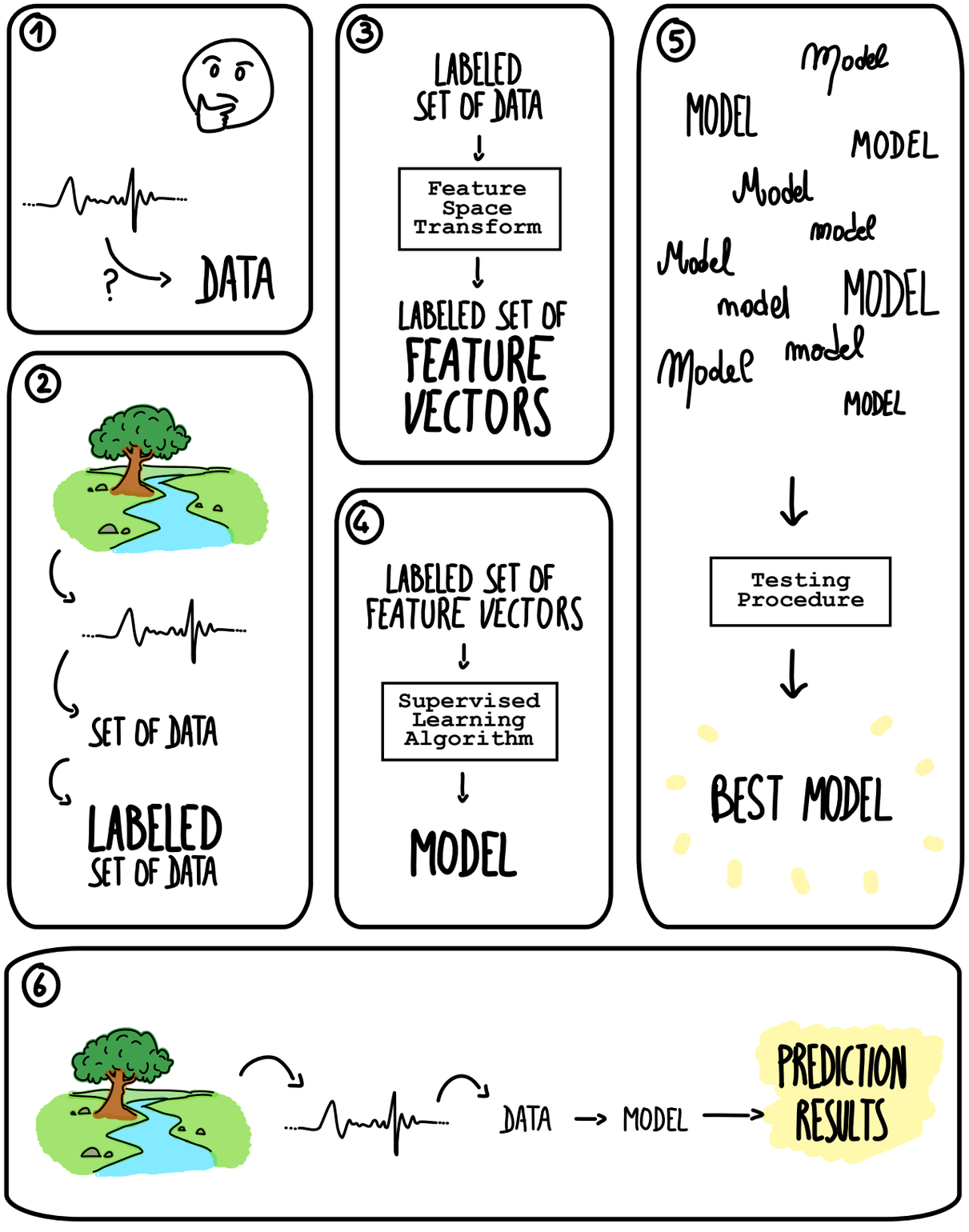
HOW TO AUTOMATICALLY ANALYZE ENVIRONMENTAL SIGNALS?



I - CONTRIBUTIONS & CONTEXT

- Presentation of the proposed architecture (detection + classification + anomaly detection),
- Machine Learning background,
- Presentation of the proposed feature extraction scheme.

1 General architecture



Data definition

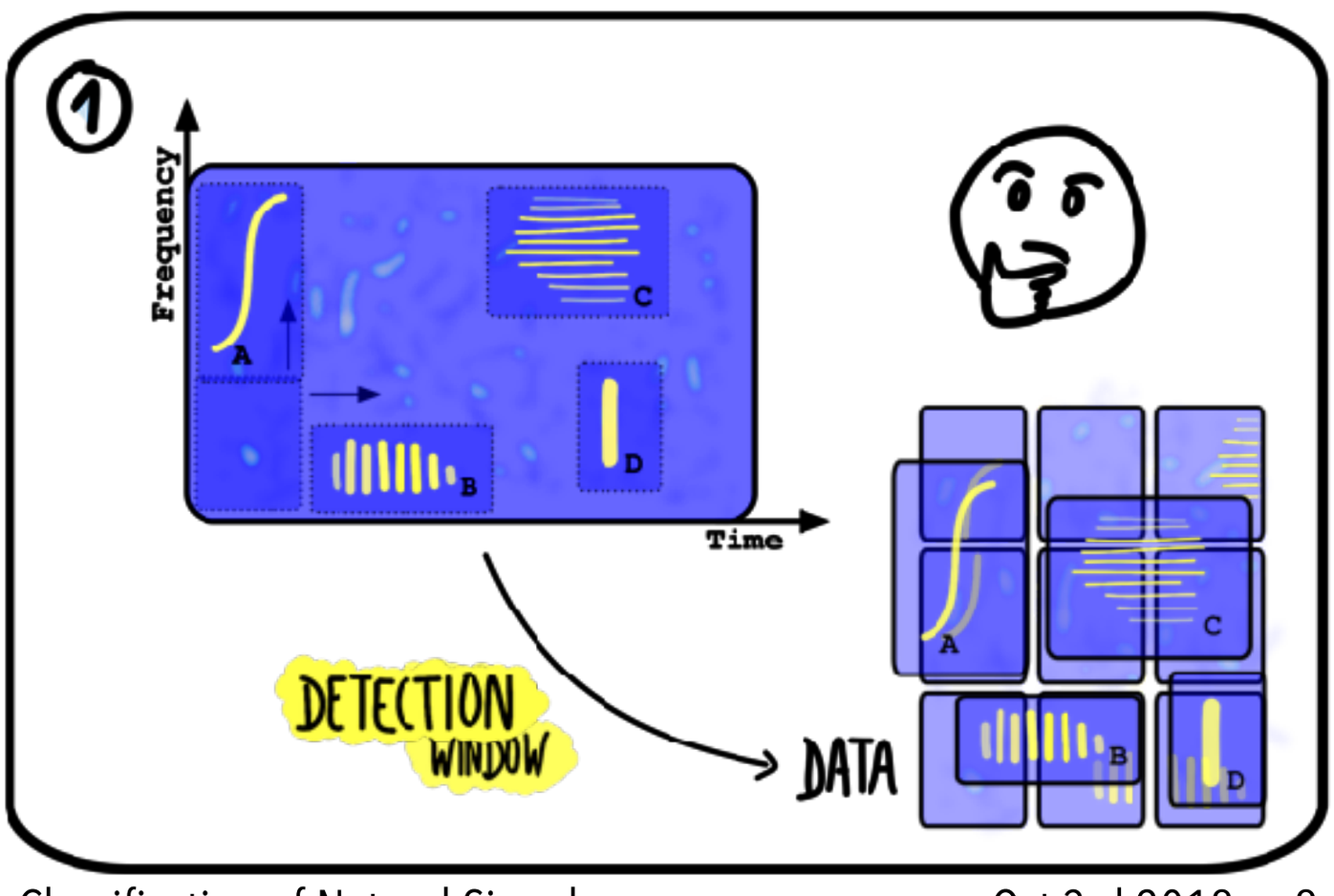
Observations are not images!

2 - Detection + Classification

Original **recording** (continuous signal): $[w_k]_{k=0}^{n_t}$

Data, **Observations** (bloc based, filtered signals):

$$[s_k]_{k=1}^{n} = h\left([w_k]_{k=n_1}^{n_1+n} \right)$$



Labeled dataset constitution

WHAT?

- Automatic classification = Automatic classification of observations into categories (**classes**),
- Set of **{ observations + labels },**
- **Label**: Category of the observation.

$$\{\mathbf{S}, \mathbf{Y}\} = \{\mathbf{s}_i, y_i\}_{i=1}^{N}$$

with $y_i \in \{0, \ldots, C-1\}$ and C the number of classes.

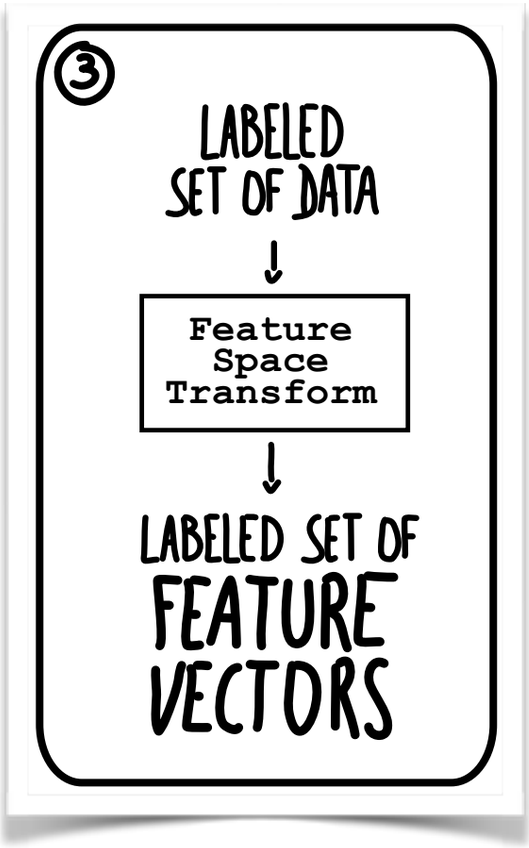
WHY?

- Supervised Machine Learning.

HOW?

- Manually,
- Collaboration with the experts,
- Transfer of human knowledge (expertise).

Features extractions



WHAT?

Labeled set of observations

$$\{\mathbf{X}, \mathbf{Y}\} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$$

- Several **low level** (not permutable) representations:
  1. Waveform, $s_k = [s_k]_{k=1}^{n}$
  2. Spectrum, $S_f = [\mathcal{F}\{s_k\}]_{f=1}^{m}$
- Generic **shape descriptors** (~ 2)

Mean, standard deviation, Skewness, Kurtosis, RMS bandwidth, Shannon entropy, Rényi entr length, rate of attack, rate of decay, threshold e rate, etc.

See manuscript page

WHY?

- Reduced dimension [D 1954],
- **Separability** of the corresponding observation classes,
- **Ordered invariant representations** [Duda&al2001].

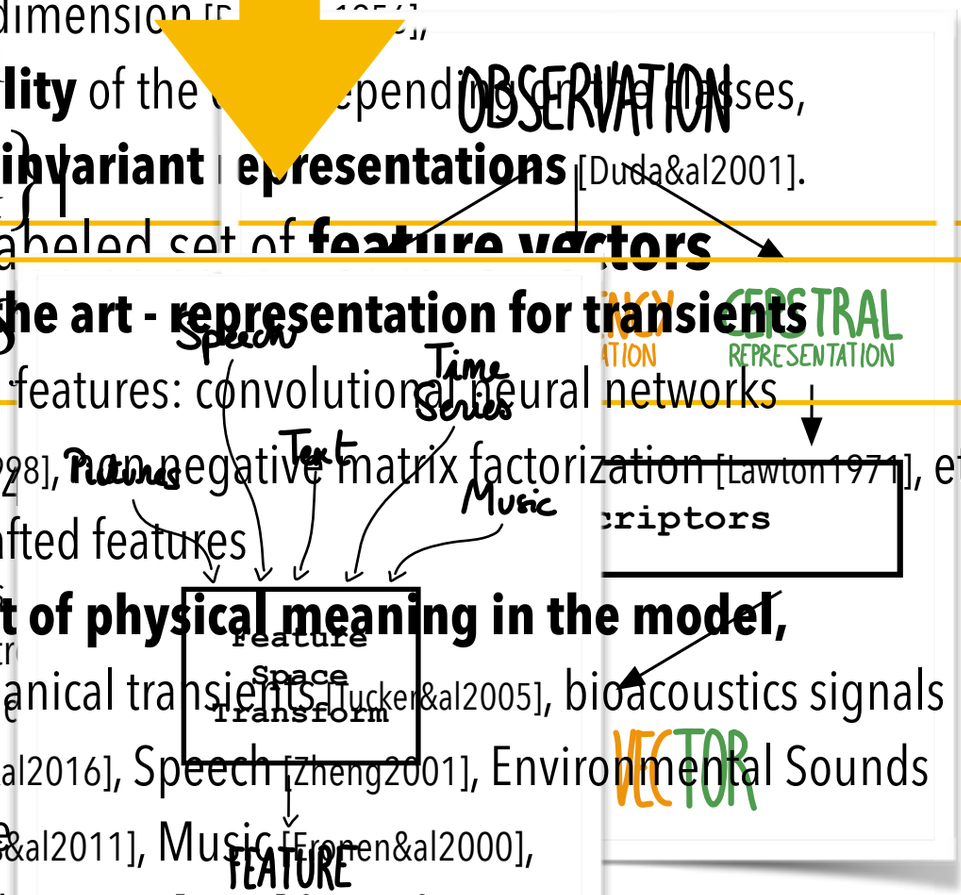
Labeled set of **feature vectors**

HOW?

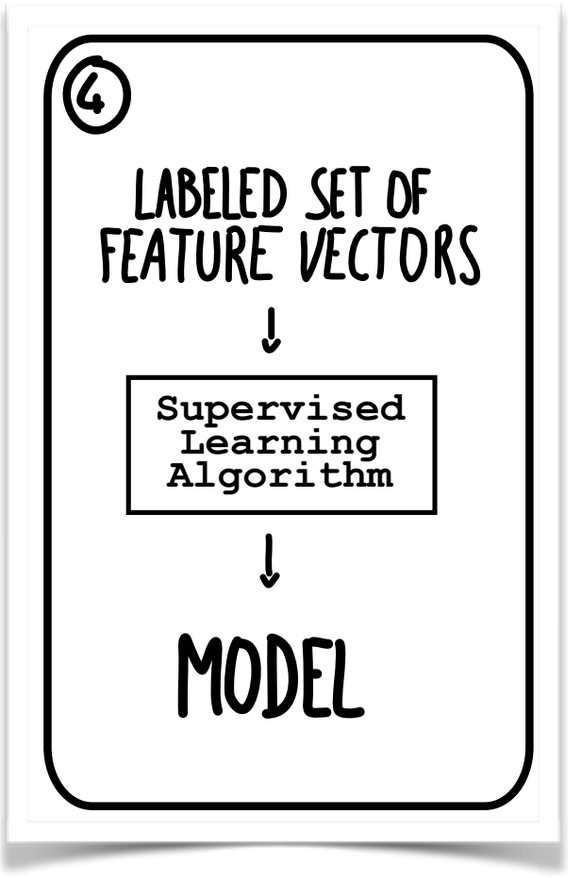
- **State of the art - representation for transients**

A. Learning features: convolutional neural networks [LeCun&al1998], non-negative matrix factorization [Lawton1971], etc.

B. Handcrafted features
   - **Input of physical meaning in the model,**
   - Mechanical transients [Tucker&al2005], bioacoustics signals [Noda&al2016], Speech [Zheng2001], Environmental Sounds [Dennis&al2011], Music [Fronen&al2000],
   - Not always **ordered invariant**.



Model training



WHAT?

- Train a **model**,
- Use of **supervised learning** algorithms,
- **Learn a separation** between the different classes from the **labeled set of observations:**

$$f : \mathbf{X} \rightarrow Y$$
$$\mathbf{X} \rightarrow y$$

HOW?

Support Vector Machine [Cortes&Vapnik95]

- Hyperplan **maximizing** the **margin**,
- **Kernel trick**: `transformation' to a space of higher dimension,
- **Soft margin.**

Random Forest [Breiman2001]

- Ensemble of decision trees [Quinlan86],
- At each split, minimize the impunity in the regions,
- Subset of features / data for each split / tree.

Model selection

WHY?

- Influence of the **observations**,
- Influence of the chosen **features**,
- Influence of the **learning algorithm**,
- Influence of the **hyperparameters.**

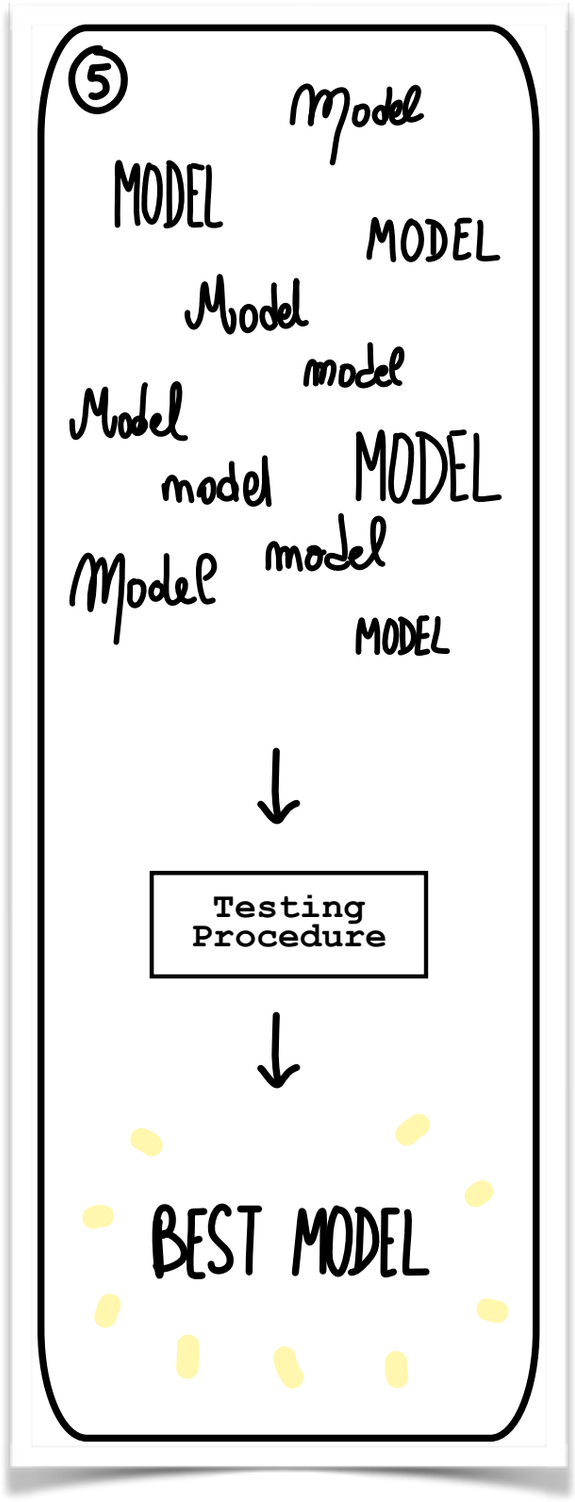
HOW?

Estimation of the model performances

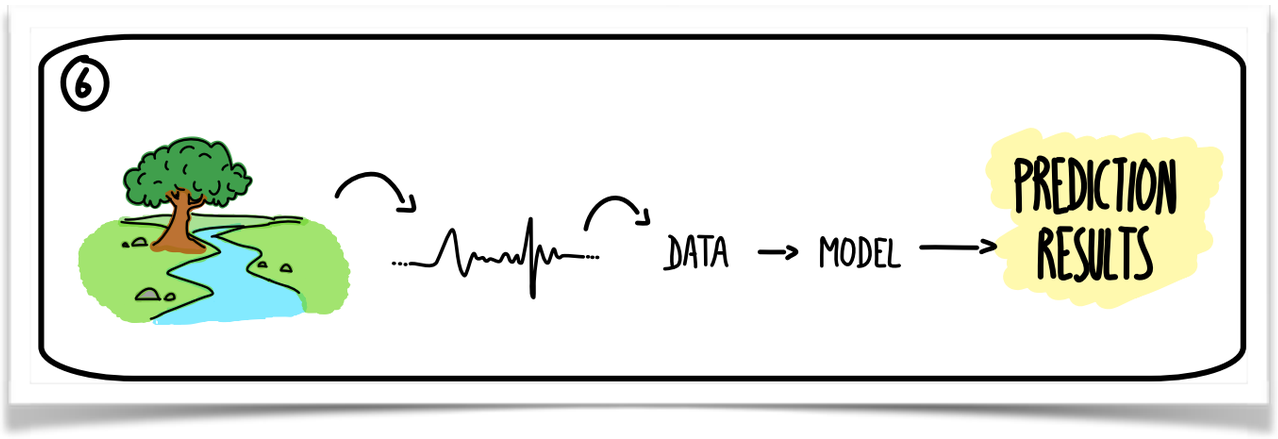
- Choice of the metric (depending on the application)

$$\textbf{Accuracy} = \frac{\#(\textbf{correct predictions})}{\#(\textbf{total predictions})}$$

- If possible [Duda&al2001]: learning - validation - test sets (Overfitting…)
- Else: **Cross-validation**
  - Learning - Testing sets (i.e., 70% - 30%)
  - Random trials of the sets (i.e., 10, 50, 200…)
  - Mean ± std results considered



Use of the model



USAGES

- **Continuous analysis**

1. **Periodic** training with the new recorded observations,
2. Analysis of the **continuous flow** of recordings.

- **A posteriori analysis**

1. Training on a given period of time,
2. Analysis of **another** period of time (evolution of the observations).

POSTPROCESSING & ANOMALY DETECTION

- **Model output**

$$p(s_i \in c) \text{ for } c \in \{0, \ldots, C-1\}$$

- **Anomaly detection**

$$c_{considered}(s_i) = \underset{c \in \{0, \ldots, C-1\}}{\text{argmax}} \; p(s_i \in c)$$

$$\texttt{if } p(s_i \in c_{considered}(s_i)) > t_c \texttt{ :}$$
$$c_{predicted}(s_i) = c_{considered}$$
$$\texttt{else :}$$
$$c_{predicted}(s_i) = \textbf{Unknown}$$

II - PRACTICAL STUDIES

1 - Volcanoes

OUTLINE

FUNDING

- Jean-Philippe Métaxian (IPGP)
- **Classification** task
- Adolfo Inza & Orlando Macedo (Arequipa observatory) - **UBINAS (Peru)**
- Ubinas & Merapi volcanoes, seismo-volcanic signals
- Francois Beauducel & Budi Dantoso (BPPTKG) - **MERAPI (Indonesia)**
- EXPERIMENT 1: Evaluation of the **model performances**
- EXPERIMENT 2: **Features** influence & selection
- EXPERIMENT 3: **A posteriori analysis of a 6 years dataset**
- 2 missions in Indonesia
- EXPERIMENT 4: Continuous analysis & **operative monitoring**
- Doctoral school - TUE , Idex, Data Institute

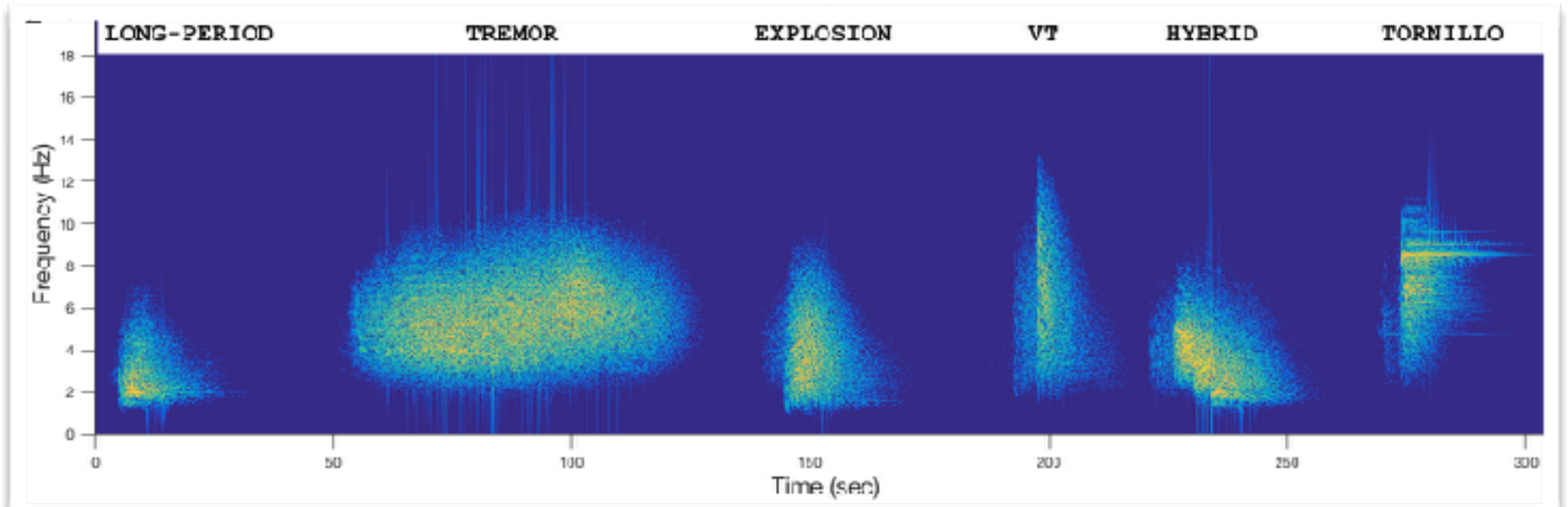
Volcano-seismic signals

UBINAS

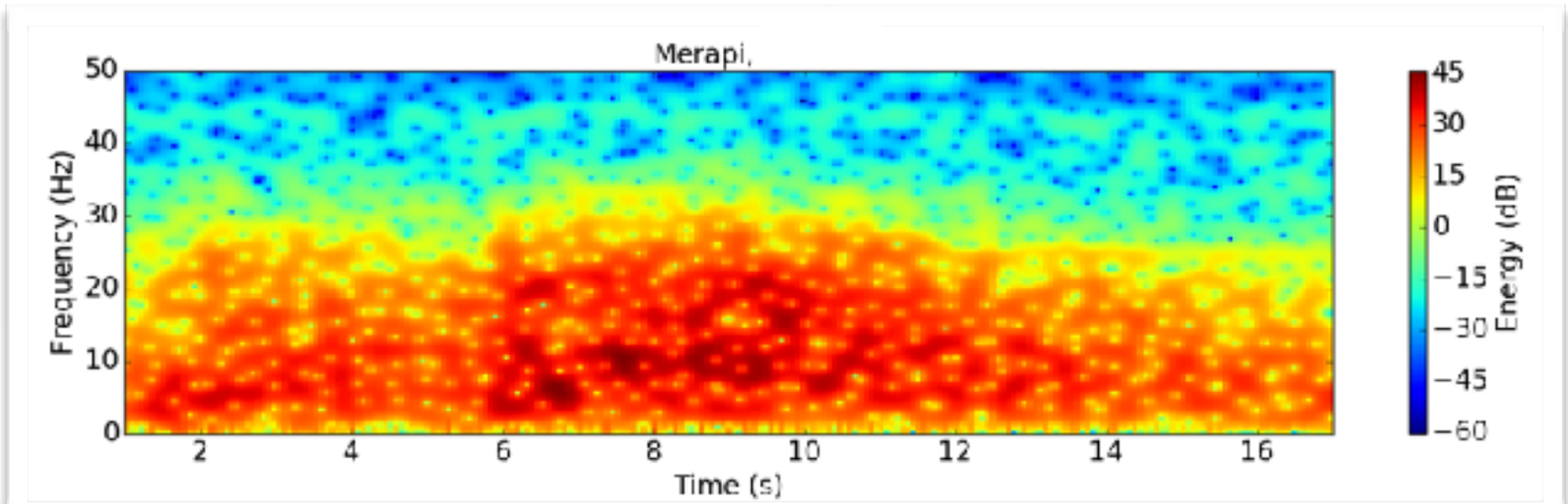
- Peruvian volcano,
- Monitored since 2006,
- **3 eruptions** since 2006.

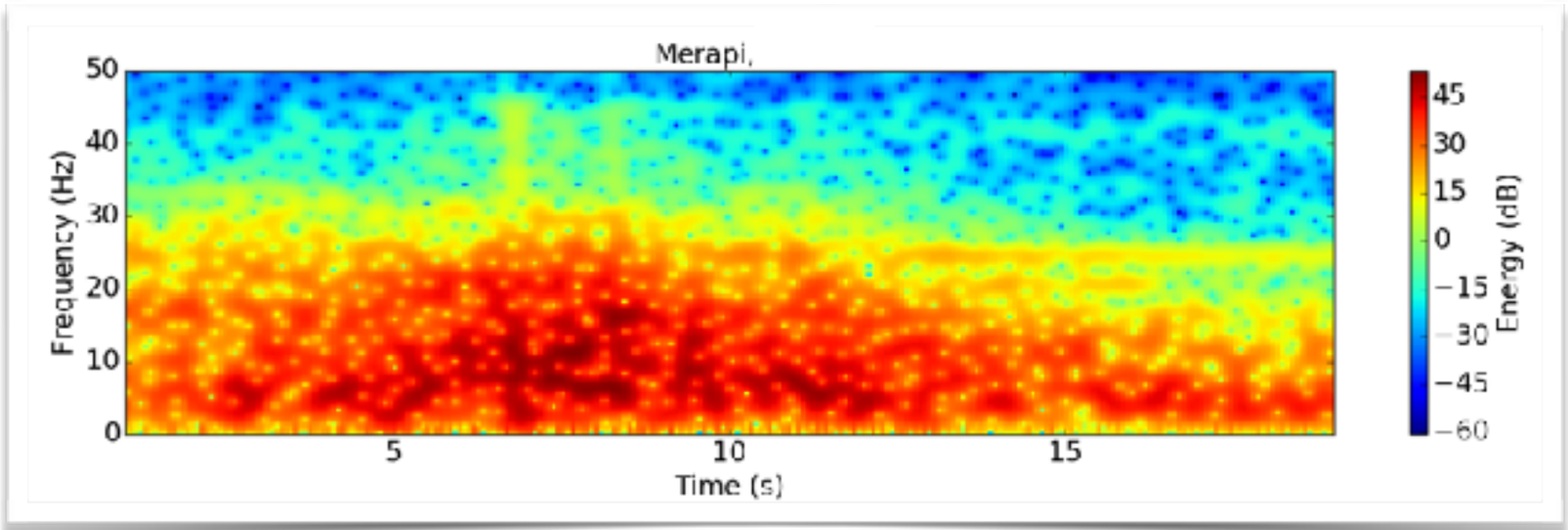
DATASET

- 109,609 labeled **observations** over **6 years** of continuous volcano-seismic recordings,
- **6 classes** (**unbalanced** dataset from **109** to **95,243** observations).



Is it difficult to classify volcano-seismic observations?

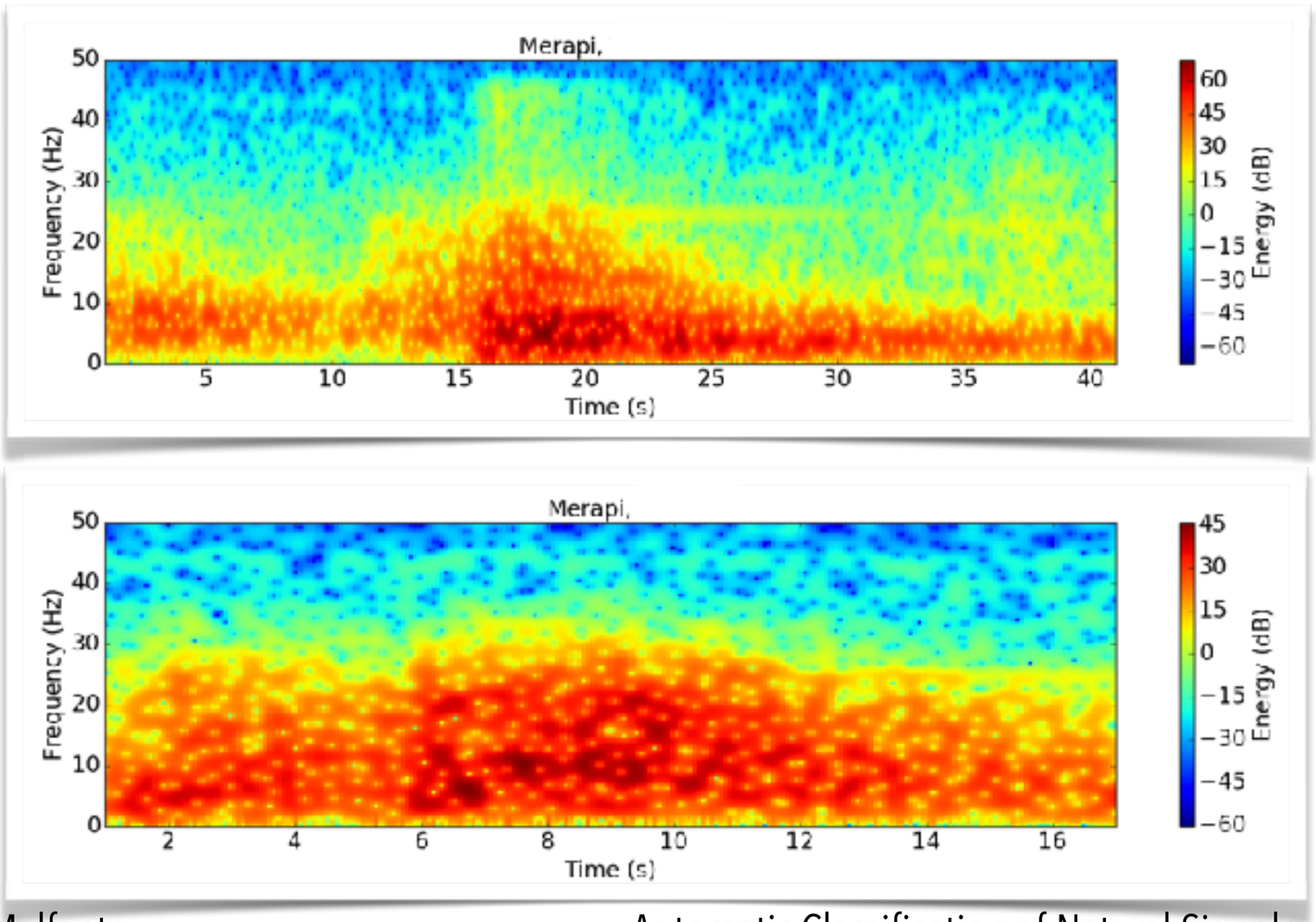




The experts said:

'Same class!'

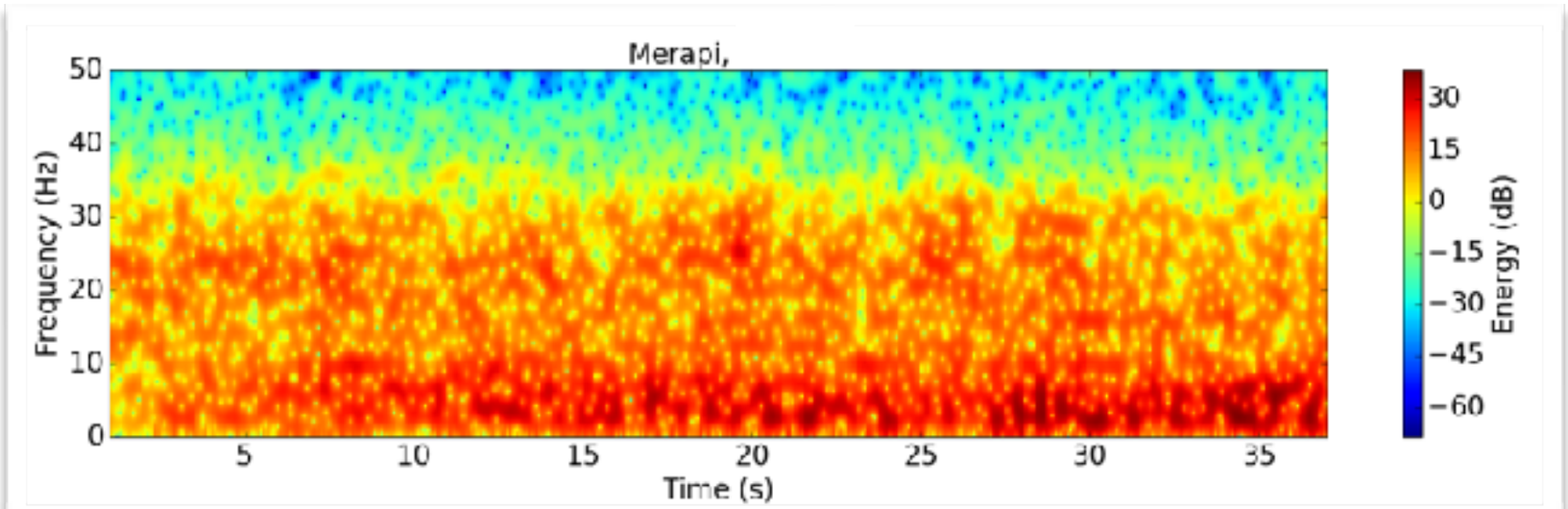
Is it difficult to classify volcano-seismic observations?

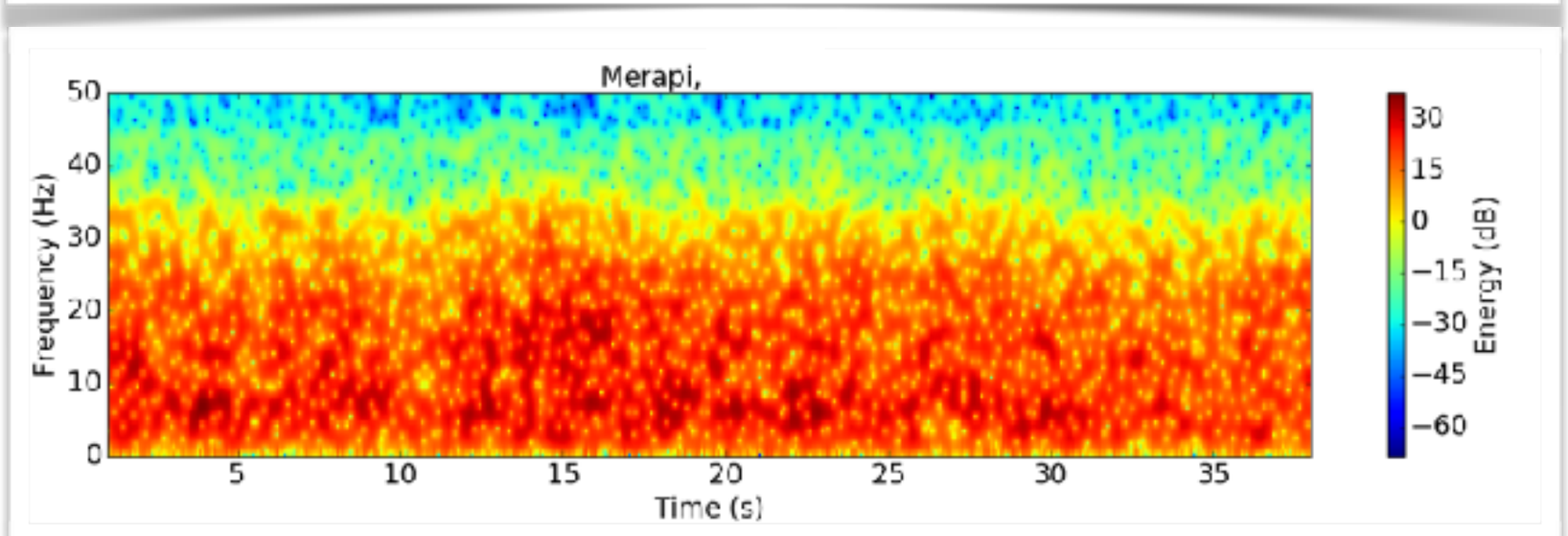


The experts said:

'Different class!'

Is it difficult to classify volcano-seismic observations?

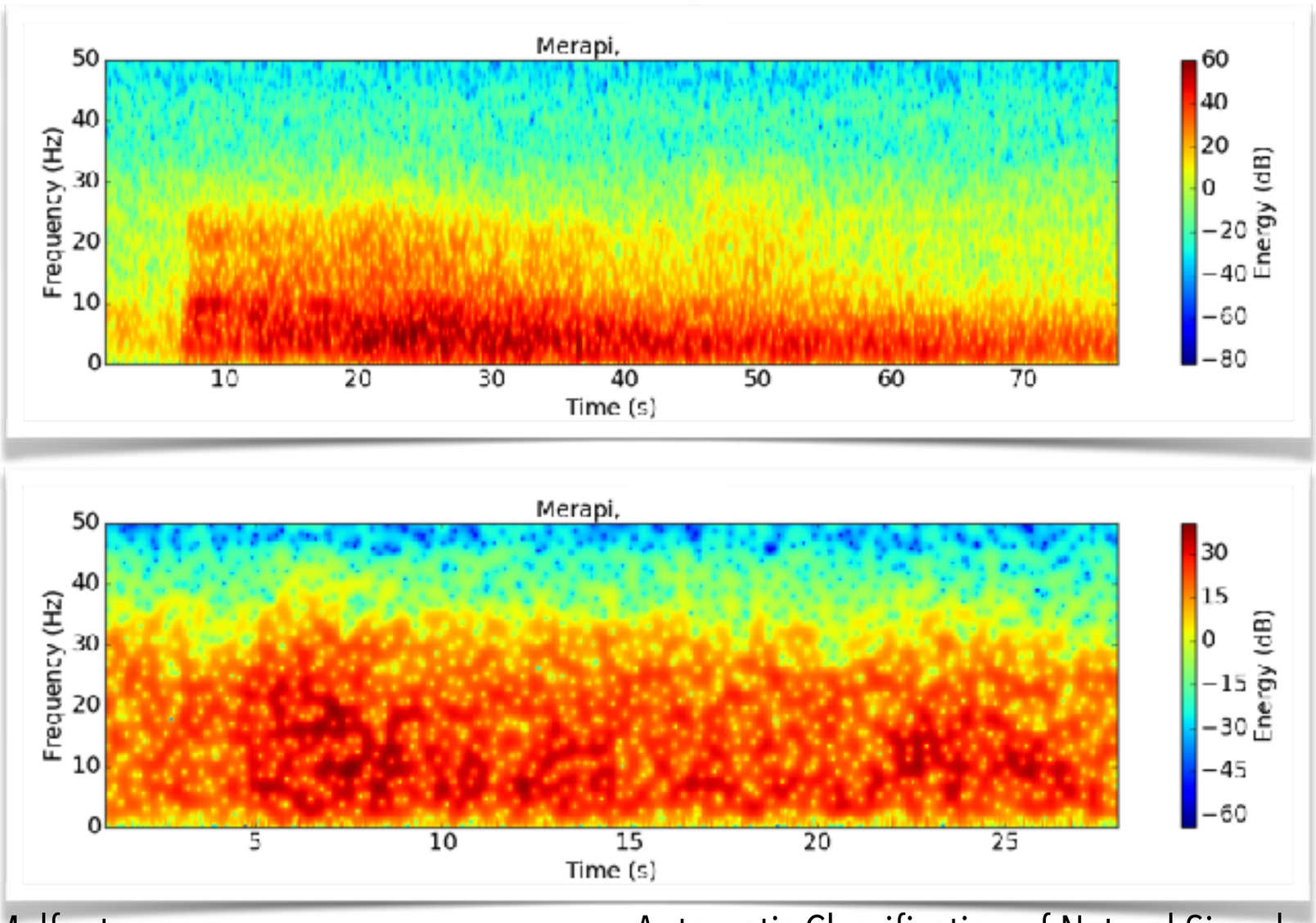




The experts said:

'Different class!'

Is it difficult to classify volcano-seismic observations?



The experts said:

'Same class!'

Evaluation of the model performances

ACCURACY
for two learning algorithms

Cross-validation, 70,856 observations, hyperparameters optimized, all features

Support Vector Machine	92.1 ± 0.54 %
Random Forest	92.5 ± 0.45 %

CONFUSION MATRIX
for the selected model

- Mixed classes have **similar causes**,
- Interest in the **accuracy** (but depends on the context).

Cross-validation, 70,856 observations, hyperparameters optimized - RF model - all features

		Expert class					
		LP	TR	VT	EXP	HYB	TOR
Predicted Class	LP	57504	457	4	1	8	-
	TR	3911	4764	-	1	3	1
	VT	372	10	487	5	12	3
	EXP	112	8	6	41	-	-
	HYB	128	6	14	1	119	-
	TOR	2	0	3	0	0	28
Accuracy		92.7%	90.8%	94.6%	84.8%	83.6%	87.6%

A posteriori analysis of 6 years of volcano-seismic recordings

HOW?

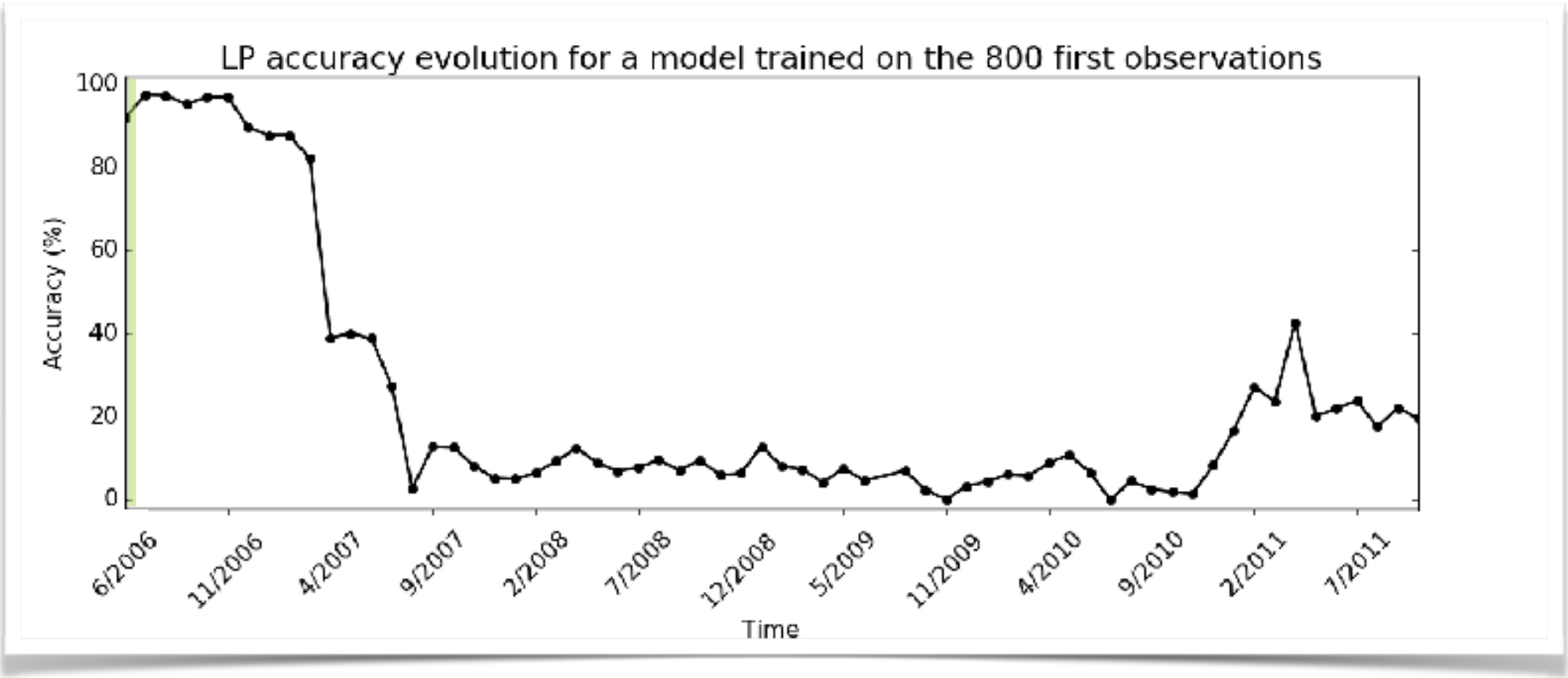
Training on the $\mathbf{N_{max} = 800}$ **first observations of each class**

- 2 weeks for LP,
- Not reached for rare classes.

WHY?

ıtion of the signals,
s of classification in the dataset.

Study the **evolution** of
tions.



II - PRACTICAL STUDIES

2 - Underwater acoustics

COLLABORATION

- Cédric Gervaise (Chair CHORUS, Fondation Grenoble INP)

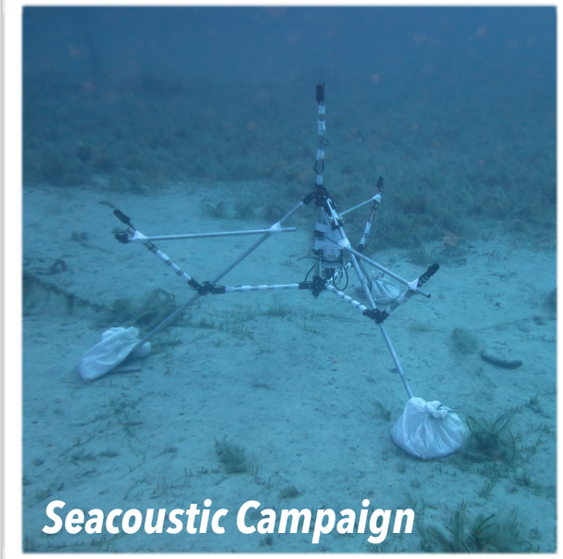
OUTLINE

- **Detection, classification + anomaly detection**
- Focus on fish sounds - SEACOUSTIC data
- ~~EXPERIMENT 1: Evaluation of the model performances~~
- EXPERIMENT 2: **Features influence & selection**
- EXPERIMENT 3: **Full analysis** of continuous recordings

Underwater acoustic data

SEACOUSTIC

- Data gathering campaign, **Passive Acoustic Monitoring** [Lossent&al2015],
- Objectives:
  (i) Determine the vitality of underwater areas,
  (ii) Evaluate the anthropogenic stress of a given area,
  (iii) Link the vitality of an area with the anthropogenic stress it faces.
- **5 areas**, several days of recording,
- STARSEO station, Pointe de la Revelatta, Corsica.


Seacoustic Campaign

Ref.	Area	Depth	Duration
1	Healthy sea-grass meadow	-20m	3.5 days
2	Healthy sea-grass meadow	-12m	1 day
3	Lower sea-grass meadow / sand border	-38m	1 day
4	Damaged meadow + Rocks	-12m	1 day
5	Rock	-12m	1 day


Seacoustic Campaign

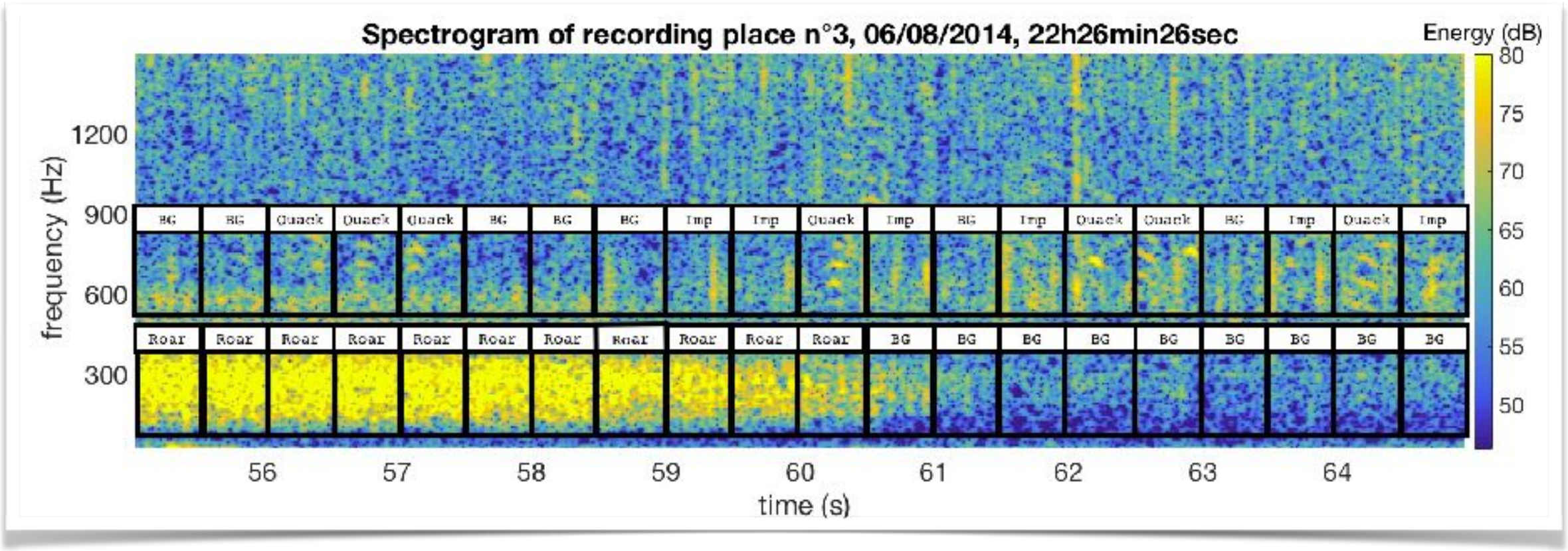
Underwater acoustic data

OBSERVATION

$\Delta_t = 0.5s$ $\Delta_f = 400Hz$

DATASET

- **913** observations (manually extracted),
- 4 **`positive'** classes (fish sounds): Quacks, Drums, Impulse, Roar.
- 1 **`negative'** class: Background.

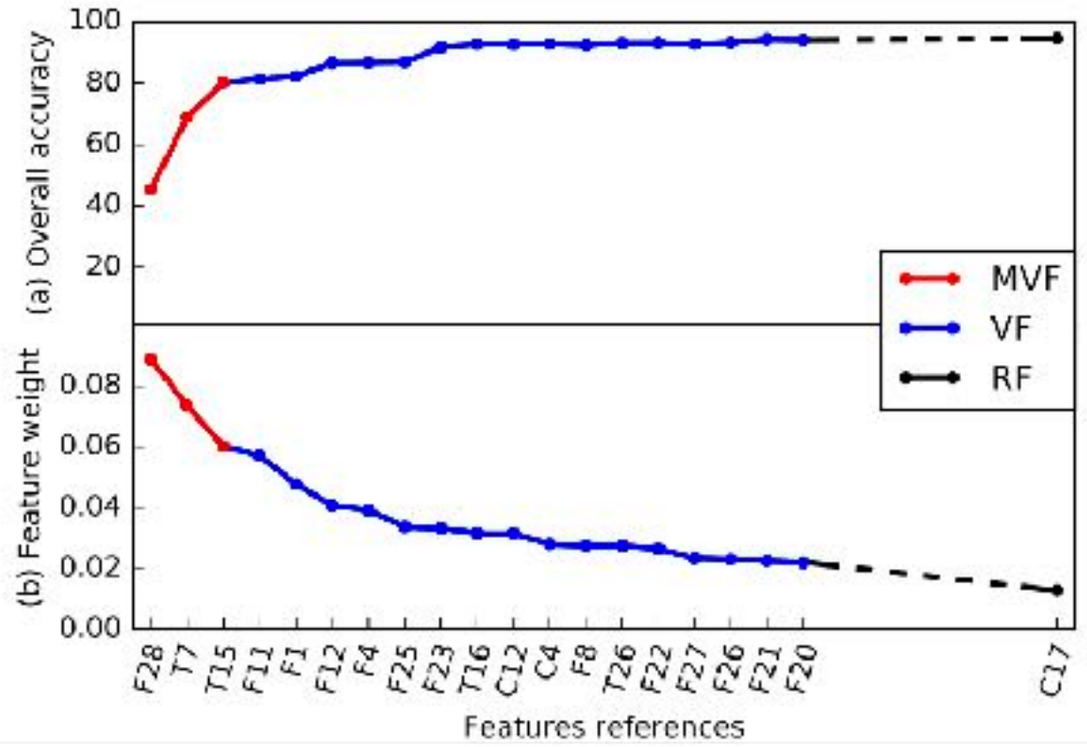


Features influence & selection

Cross-validation (30% - 70%) - RF - 200 trees - entropy

	Accuracy
Time (40)	90.1 ± 2.0 %
Frequency (40)	91.1 ± 2.7 %
Cepstral (40)	91.4 ± 3.0 %
All features (120)	96.9 ± 2.0 %
Most Valuable Features (3)	81.3 ± 0.85 %
Valuable Features (19)	95.6 ± 0.79 %
MFCC (26)	72.5 ± 3.3 %
Deep features	75.0 ± 1.99 % to 89.5 ±1.15 %

Use of RF for feature selection

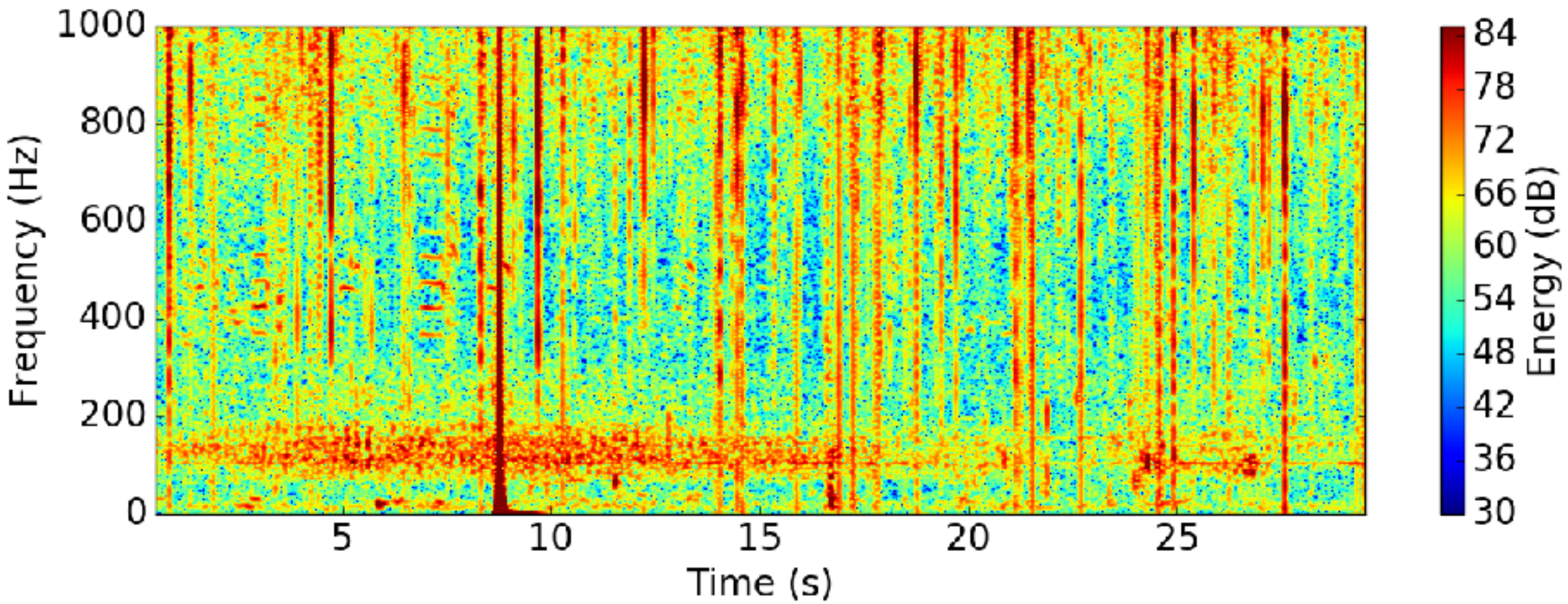


Three Most Valuable Features

1. Energy kurtosis (from spectrum)
2. Mean kurtosis (from waveform)
3. Threshold crossing rate (from waveform)

$$\frac{1}{\#(ThresholdCrossing)}\sqrt{\frac{\sum_i (E_i - E_{it})^4 \cdot E_t}{E_t \cdot RMS_E^4}}$$

Full analysis of continuous recordings



Full analysis of continuous recordings

1 Check the model performances under **various using conditions**

Learning **Area#3 - given time**

Testing 1 - Area#3 - **Different time**

Overall accuracy = 93.4 %

		True Class (ground truth)					
		B	D	I	Q	R	U
Predicted Class	B	969	2	2	-	2	23
	D	-	133	1	-	-	2
	I	-	-	208	-	-	-
	Q	5	-	-	245	-	2
	R	-	-	-	-	58	1
	U	39	16	11	40	13	624
Accuracy		95.7%	88.1%	93.7%	86.0%	79.5%	95.7%

SVM model, all features

Testing 2 - **Area#2** - **Different time**

Overall accuracy = 80.9 %

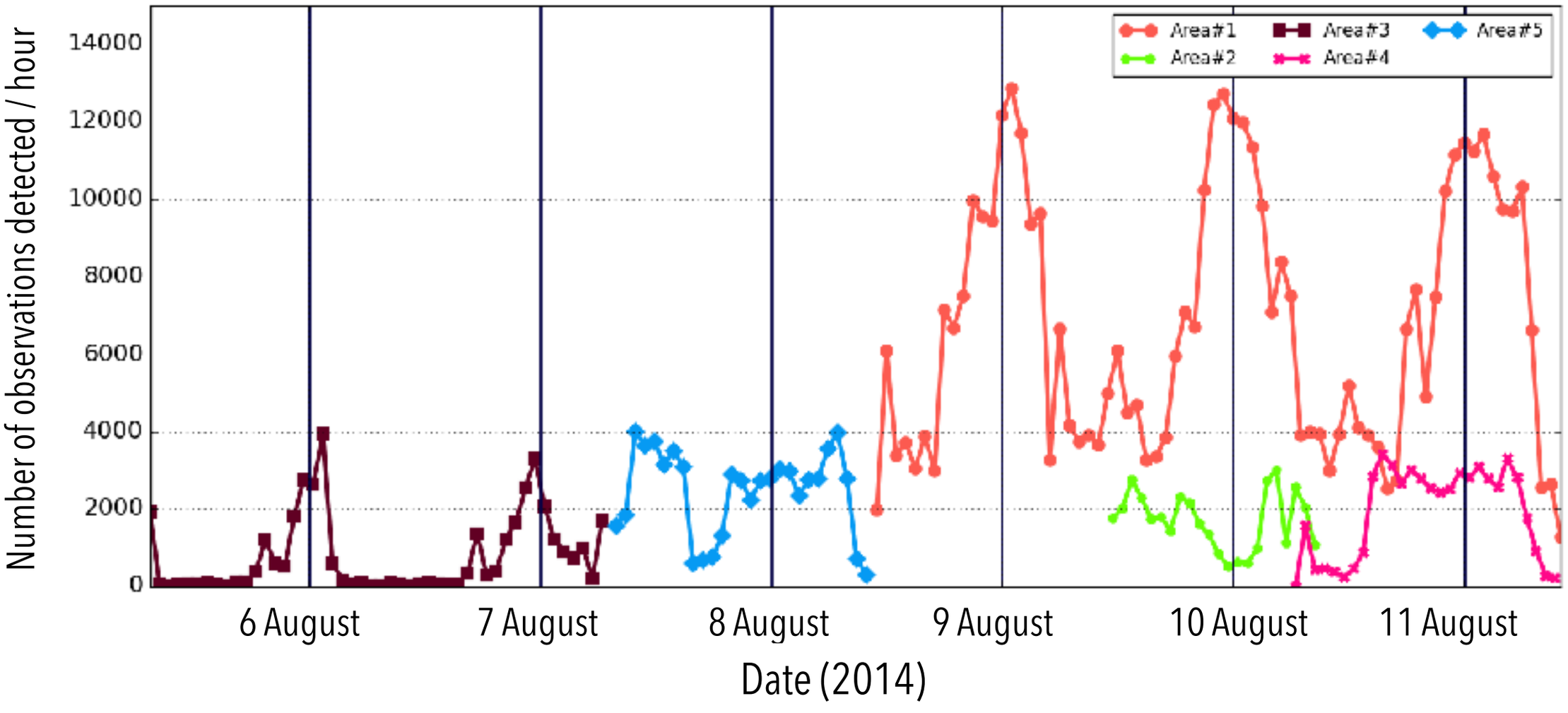
		True Class (ground truth)					
		B	D	I	Q	R	U
Predicted Class	B	885	6	4	-	-	21
	D	3	40	-	-	-	8
	I	16	-	21	-	-	-
	Q	6	-	-	256	-	-
	R	-	-	-	-	-	-
	U	207	24	15	127	-	651
Accuracy		79.2%	57.1%	52.5%	66.8%	-	95.7%

SVM model, all features

Full analysis of continuous recordings

2 Analysis of the 5 different areas

Hourly number of observations detected as 'Drums' (various recording areas)



Conclusion

Summary

About machine learning & applications

- Scientifically mature tools?
- Usually for the same data: images, text, speech
- **What can we do for `unusual' data?**

Focus on the automatic processing of environmental signals.

- Different applications, different constraints, but some **common points:**

1. **Keep things as general as possible,**
2. **Formalize some aspects,**
3. **Propose operational tools (real world data),**
4. **Knowledge on the environment,**
5. **Transfer of human knowledge (expertise), input of physics in the models.**

Prospects

6	MORE MODELS
3	POST-PROCESSING
4	**MULTI-*** ANALYSIS
5	LABELING CONSTRAINT

Interest for other types of analysis:
1. Unsupervised analysis (less biased by human knowledge),
2. Mapping (for online visual monitoring).

Idea of hybrids architectures, running several types of analysis in parallel.

Apply **time regularization** methods (temporal coherence).

Comparison of volcanoes

Multi **time scales,** Multi**source,** Multi**modality,** Multi**label.**
Hierarchical models.

Exploration of unknown environment in general.

To be reduced. **Semi-supervised** analysis.

PhD, coding but also

International journal publications
Signal Processing Magazine, **Journal of Acoustical Society of America,** **Journal of Geophysical Research.**

Conferences
ASA 2016, Salt Lake City (USA) + Best student paper award + Lay language paper **EGU 2017, Vienna (Austria)** **EUSIPCO 2017, Kos (Greece)** **GRETSI 2017, Juan-les-pins (France)** **OCEANS 2018, Kobe (Japan)**

PhD, coding but also

Seminars

- Automatic Classification of Underwater Acoustics Data (poster), Machine Learning Summer School 2016, Cadiz (Spain)
- Automatic Classification and Detection of Fish Sounds (poster), *Deep Learning Summer School 2016*, Montreal (Canada)
- Machine Learning & Bioacoustics, *Serenade 2016*, Brest (France)
- Bandung (Indonesia),
- BPPTKG Yogyakarta (Indonesia),
- IPGP Paris (France)

Teaching

- Label RES (Recherche Enseignement Supérieur) - 128h of teaching + 80h of training (signal processing, digital signal processing, C, java)
- Supervised Machine Learning & Applications to the Automatic Classification of Fish Sounds, *Chorus Workshop 2016*, Grenoble (France)
- Machine Learning: Introduction on general concepts & focus on supervised algorithms, training for BEST european student association, Grenoble (France)

Thank you

Questions, interrogations, inquiries, comments, arguments, debate, discussions, dialogue, observations, opinions, reflections, statements, analysis, exchange, thoughts, remarks, …

go for it!