Philippe Lenca
Ronan Fablet
(Telecom Bretagne)

"Méthodes de fouille et d'analyse"

26 Novembre 2013, Ifremer, Brest

Ifremer

Lab-STICC

http://wwz.ifremer.fr/bigdata

Méthodes de fouille et d'analyse

Philippe Lenca & Ronan Fablet
Institut Telecom, Telecom Bretagne

UMR CNRS 6285 Lab-STICC
DECision aId and knowleDge discovEry project
Traitements, Observation et Méthodes Statistiques
Université européenne de Bretagne
[prenom.nom]@telecom-bretagne.eu

TIC et mer: nouveaux défis et solutions.
26 Novembre 2013, Brest, France

Plan

Plan

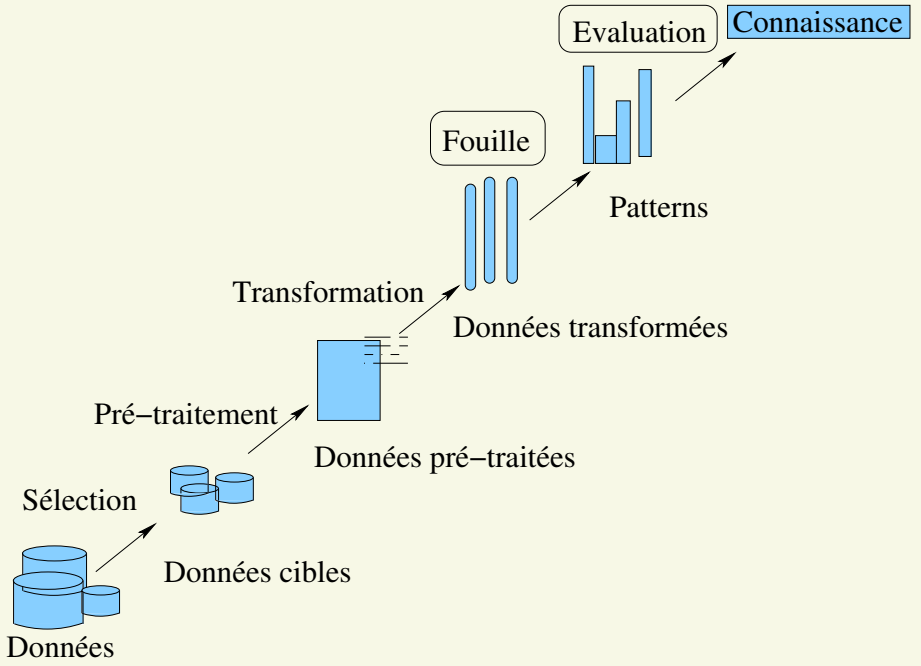
Quelques définitions ...

- processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données [PCKW89]
- processus complexe permettant l'identification, au sein des données, de motifs valides, nouveaux, potentiellement intéressants et les plus compréhensibles possible [FPSSU96]
- processus interactif et itératif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par l'utilisateur-analyste qui y joue un rôle central [KNZ01]

↪ Processus, données, motifs exploitables, interaction.

Un processus en plusieurs étapes

Linéaire

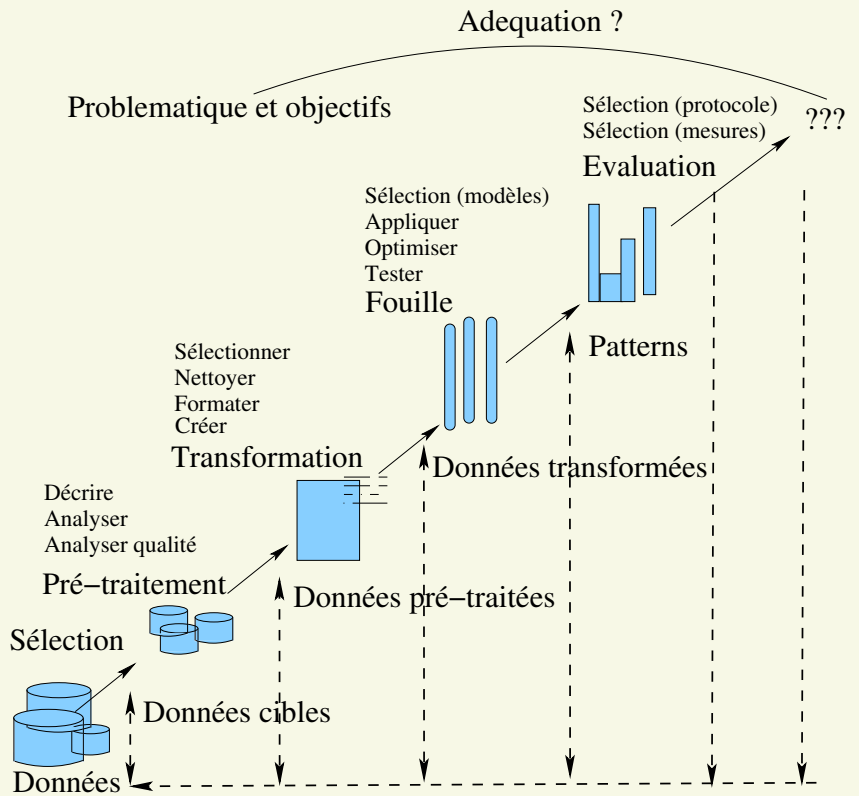


↪ . . . le rêve

TELECOM Bretagne

Un processus en plusieurs étapes

Mais un fleuve pas si tranquille (contextualisation du processus)



↪ Nombreuses itérations (choix multiples, optimisation), mais on avait le temps.

Fouille : algorithmes & tâches de fouille

Top 10 algorithmes [WKQ+08]

1. C4.5 (apprentissage supervisé, 1993)
2. K-Means (classification, 1967)
3. SVM (apprentissage statistique, 1995)
4. Apriori (recherche d'associations, 1994)
5. EM (apprentissage statistique, 1977)
6. PageRank (fouille de liens, 1998)
7. AdaBoost (bagging et boosting, 1997)
7. kNN (apprentissage supervisé, 1967)
7. Naive Bayes (apprentissage supervisé, 1763)
10. CART (apprentissage supervisé, 1984)

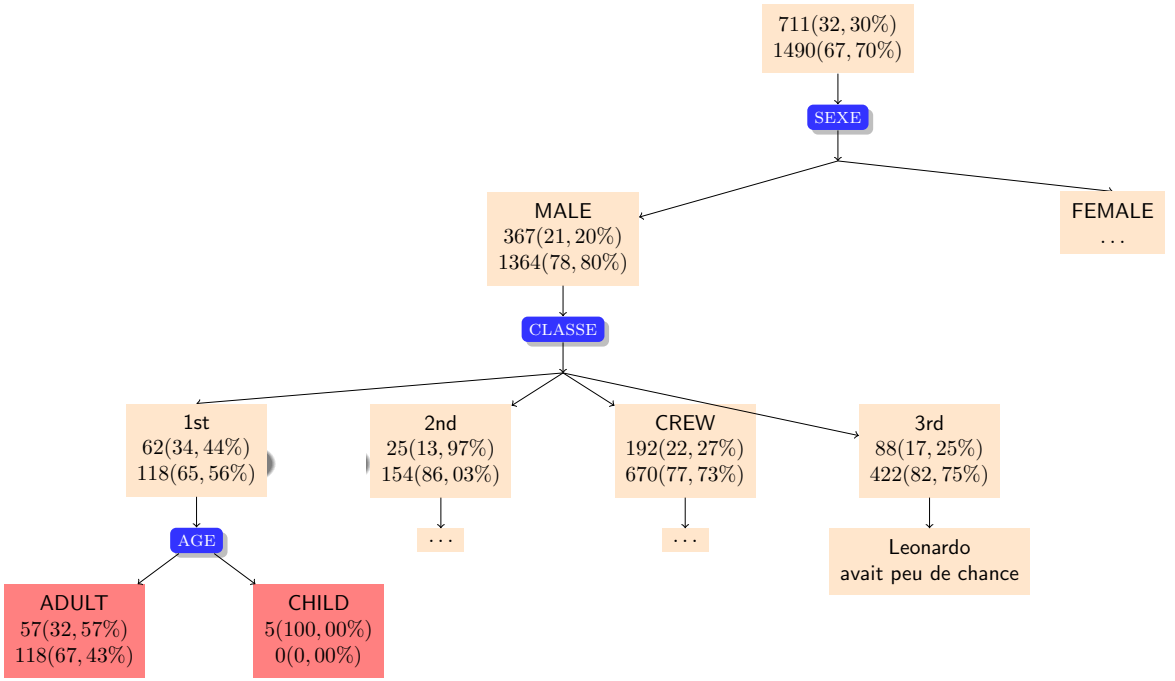
Grandes familles de tâches

1. prévision
2. classification
3. classement
4. recherche de co-occurences

↪ Apprentissage supervisé vs. non supervisé.

Exemple : arbres de décision

Prédire les valeurs prises par une variable cible à partir d'un ensemble de variables explicatives.



↪ C4.5 [Qui93], CART [BFOS84].

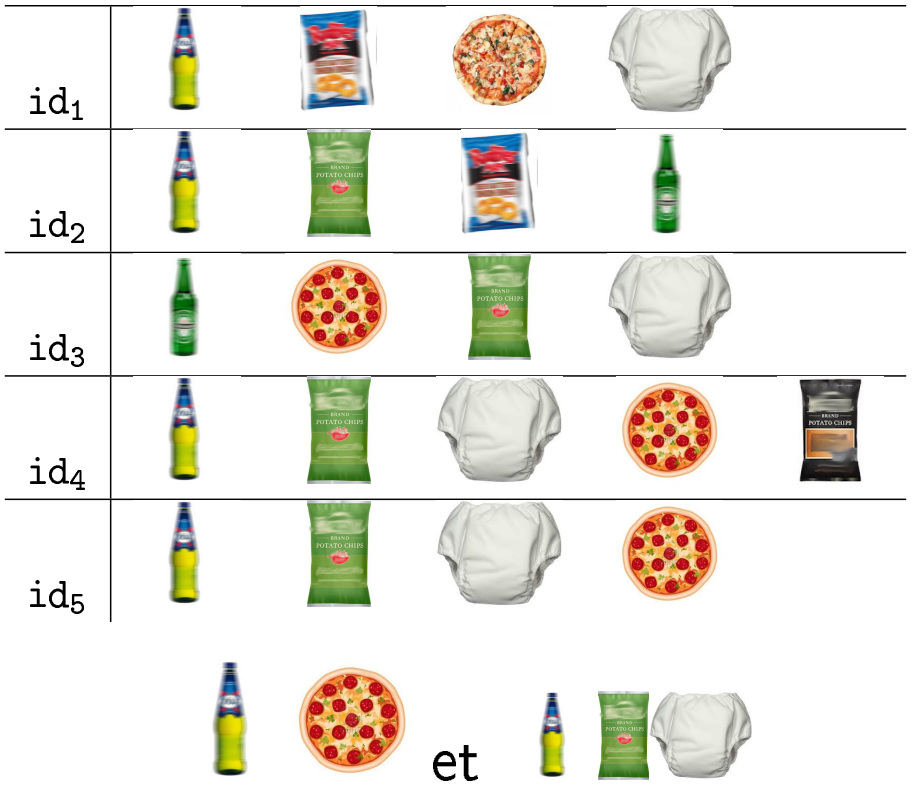
Exemple : classification

Diviser les données de telle façon à ce qui se ressemble s'assemble au sein d'un même groupe.

↪ K-Means [Ste57, Mac67].

Exemple : recherche d'associations

Recherche de co-occurences dans les données.



↪ Apriori [AS94].

Plan

Flots de données

Flot de données

- séquence ordonnée de données arrivant de façon continue, en nombre illimité, avec une grande rapidité, et dont la distribution change avec le temps.
- exemple typique : données issues de capteurs.

↪ Le traitement des flots par des systèmes (limités en mémoire, cpu et stockage) nécessite des algorithmes ne lisant les données qu'une seule fois (ou un nombre de fois très limité).

Flots de données

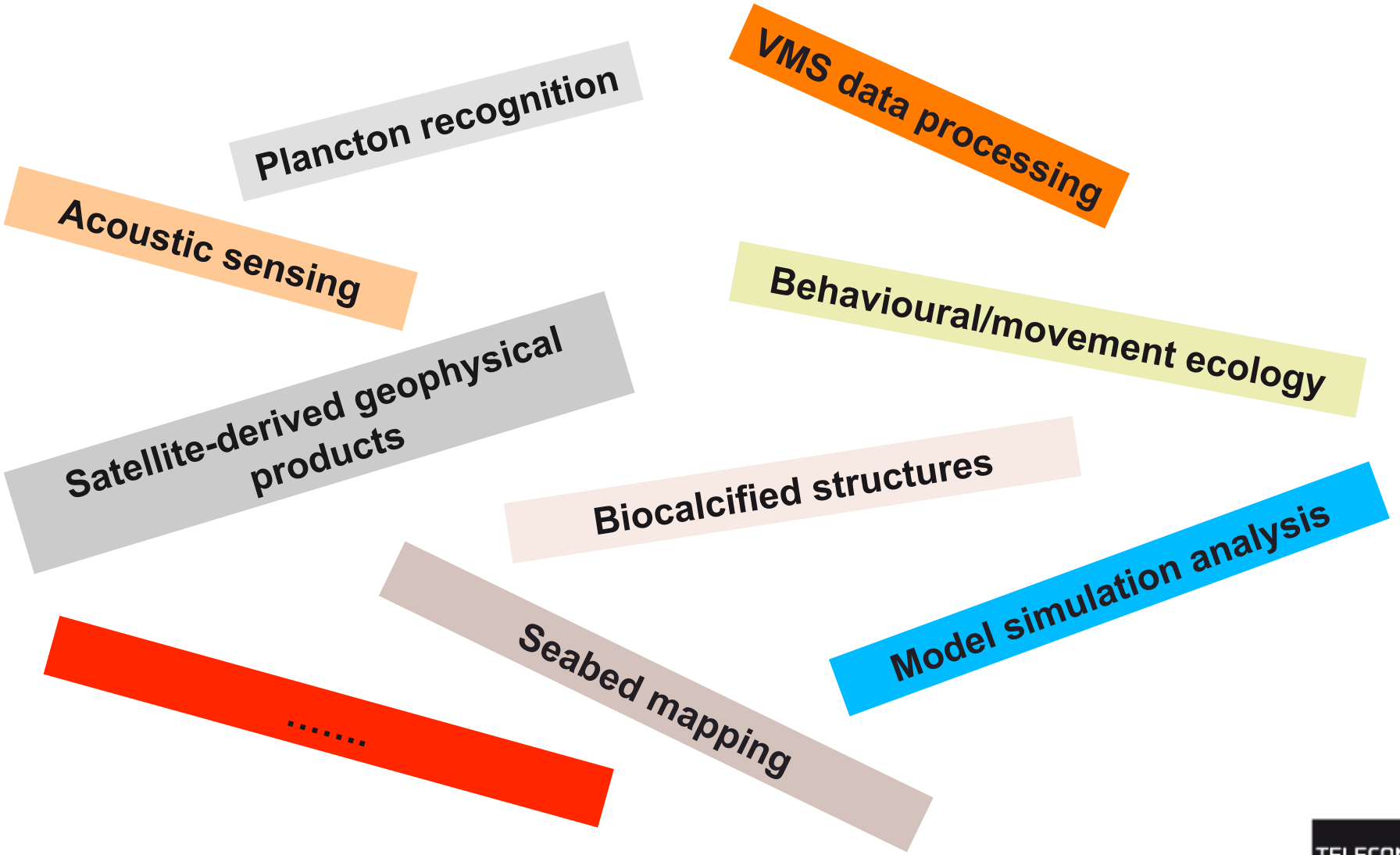
Différences entre données classiques et flots [Gam12]

	Classique	Flot
Taille	finie	flot continu
Hypothèse	i.i.d	non-i.i.d.
Évolution	statique	non-stationnaire
Ordre	indépendant	dépendant
Accès	aléatoire	séquentiel
Lecture	multiple	unique
Temps de traitement	*illimité*	restreint
Mémoire disponible	illimité	fixe
Distribution	non (oui)	oui
Construction	batch	incrémentale
Stabilité	statique	évolutive
Résultat	précis	approché

↪ Différentes étapes du processus vont être de plus en plus confondues. L'évaluation pose également de nouveaux problèmes.

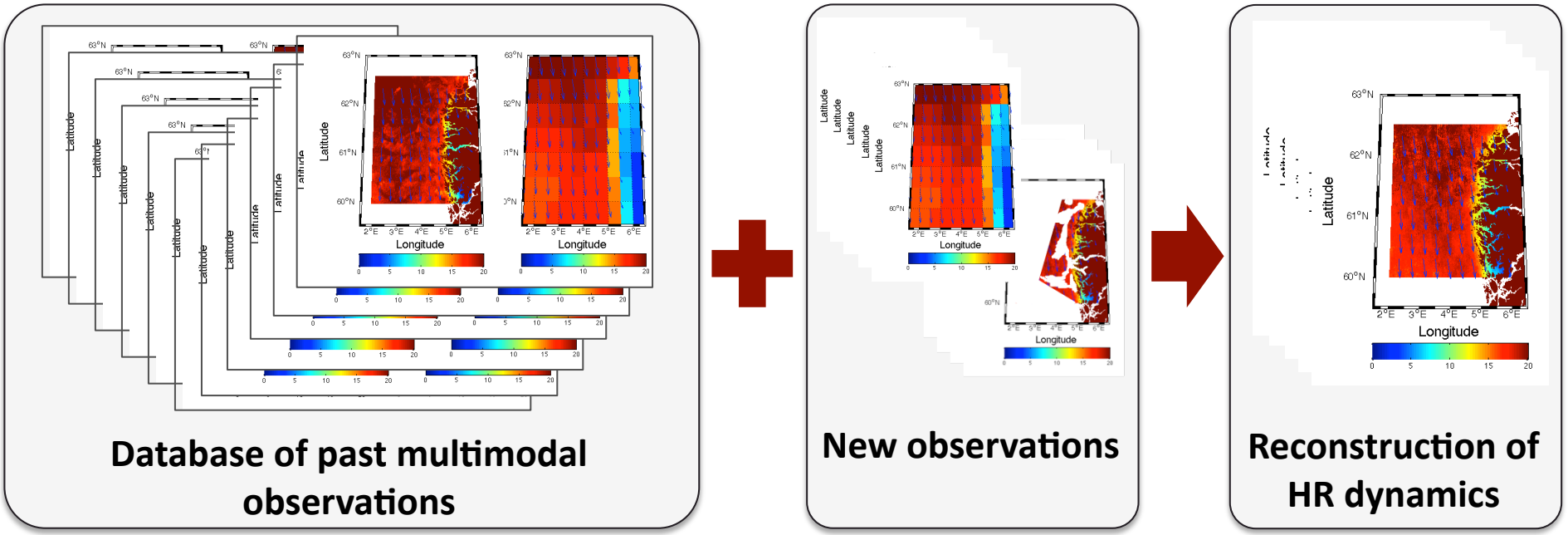
Plan

Des applications en océanographie et écologie marine



Un exemple d'application en mode supervisé

- ***Reconstruction de champs de vent haute-résolution***

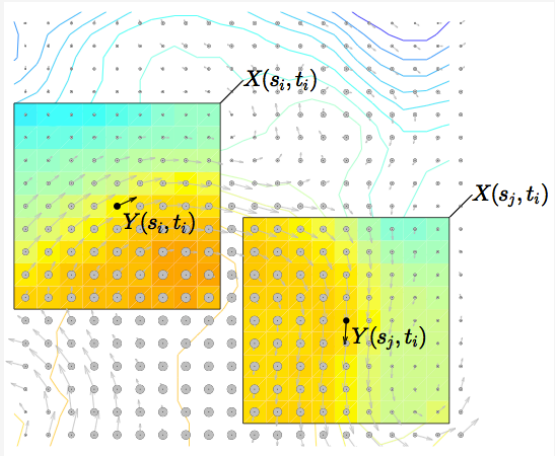


Approche : Apprentissage supervisé d'une fonction de transfert entre « vents ECMWF » et « vents SAR »

He-Guelton et al, 2013

Un exemple d'application en mode non-supervisé

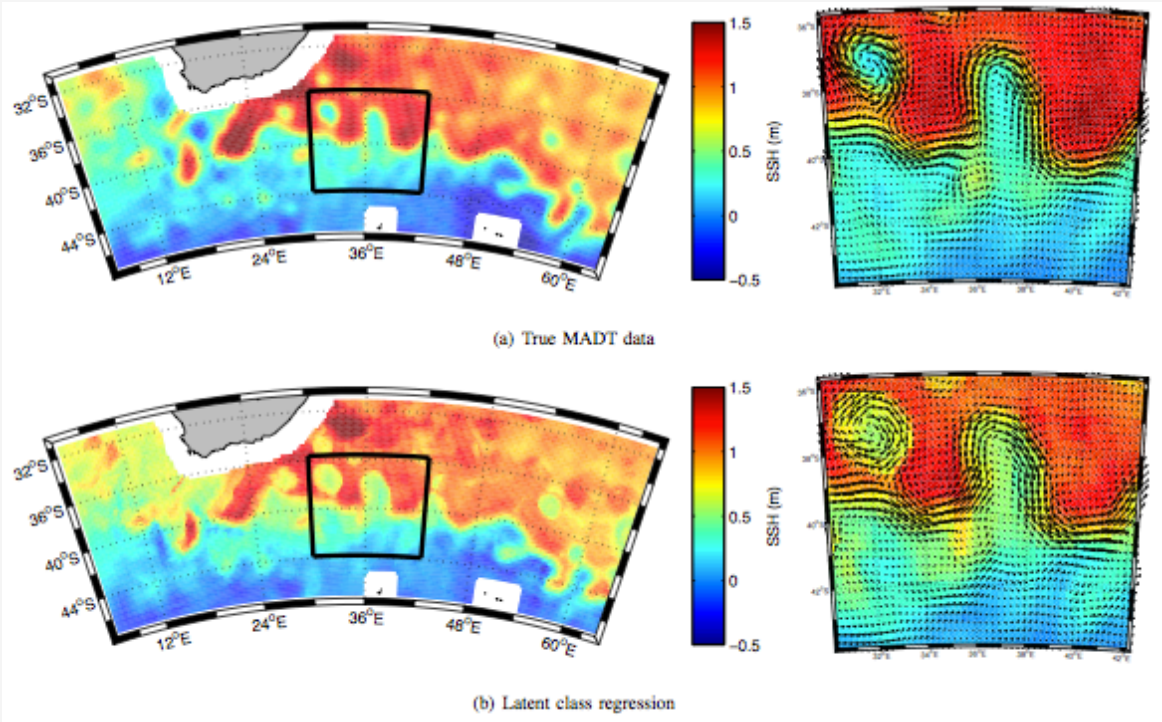
- **Relation courants altimétriques / température de surface ?**



Théorie SQG : relation locale entre les variations locales de température et le courant

Apprentissage à partir d'observations conjointes

Application to a one-year AMSR/SST-AVISO/SSH series



(a) True MADT data

(b) Latent class regression

Analyse non-supervisée pour identifier et suivre les modes dynamiques en jeu

Tandeo et al, 2013

Plan

References I

[AS94] Rakesh Agrawal and Ramakrishnan Srikant.
Fast algorithms for mining association rules in large databases.
In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *VLDB*, pages 487–499. Morgan Kaufmann, 1994.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C.J. Stone.
Classification and Regression Trees.
Wadsworth International, 1984.

[FPSSU96] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors.
Advances in Knowledge Discovery and Data Mining.
AAAI/MIT Press, 1996.

[Gam12] João Gama.
A survey on learning from data streams: current and future trends.
Progress in AI, 1(1):45–55, 2012.

[KNZ01] Y. Kodratoff, A. Napoli, and D. Zighed.
Bulletin de l'association fran caise d'intelligence artificielle, extraction de connaissances dans des bases de données, 2001.

[Mac67] J. B. MacQueen.
Some methods for classification and analysis of multivariate observations.
In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, page 281–297. University of California Press, 1967.

References II

[PCKW89] K Parsaye, M. Chignell, S. Khoshafian, and H. Wong.
Intelligent Databases; Object-Oriented, Deductive Hypermedia Technologies.
John Wiley & Sons, 1989.

[Qui93] J. Ross Quinlan.
C4.5: Programs for Machine Learning.
Morgan Kaufmann, 1993.

[Ste57] H. Steinhaus.
Sur la division des corps matériels en parties.
Bull. Acad. Polon. Sci., 4(12):801–804, 1957.

[WKQ+08] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus F. M. Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg.
Top 10 algorithms in data mining.
Knowl. Inf. Syst., 14(1):1–37, 2008.