

Extraction de motifs spatio-temporels: co-localisations, séquences et graphes dynamiques attribués

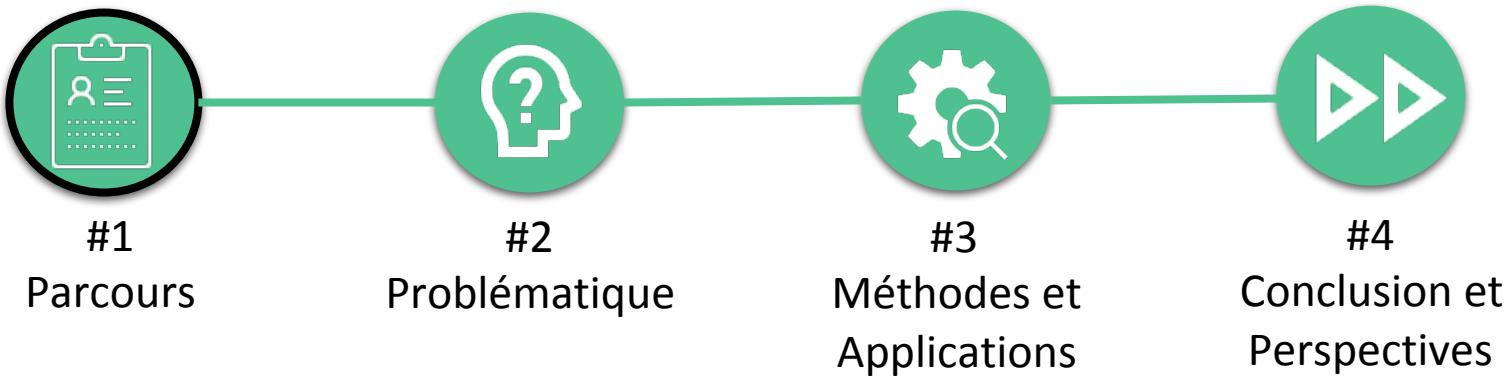
Frédéric Flouvat

Soutenance d'Habilitation à Diriger des Recherches en informatique
16 octobre 2019

Membres du jury:

Stéphane Bressan	Professeur associé, National University of Singapore, Singapore	Rapporteur
Philippe Fournier-Viger	Professeur, Harbin Institute of Technology, China	Rapporteur
Alexandre Termier	Professeur, Université Rennes 1, France	Rapporteur
Silvère Bonnabel	Professeur, Université de la Nouvelle-Calédonie	Examinateur
Jean Diatta	Professeur, Université de La Réunion, France	Examinateur
Nazha Selmaoui-Folcher	Maître de conférences HDR, Université de la Nouvelle-Calédonie	Garant HDR





Parcours Professionnel

Rattaché à l'Institut des Sciences Exactes et Appliquées (EA 7484 ISEA)
Axe "Complexité et science des données"

Depuis 2017

Rattaché au Pôle Pluridisciplinaire de la Matière et de l'Environnement (EA3325 PPME)
Equipe "Extraction et gestion des connaissances" dirigée par N. Selmaoui-Folcher

Depuis 2012

Membre du Labex Corail
Les récifs coralliens face au changement climatique

ATER en informatique
Université Lyon 1 / LIRIS

2007-2008

Depuis 2008

Maître de conférences en informatique
Université de la Nouvelle-Calédonie

Thèse en informatique
2003-2006
Vers des solutions adaptatives et génériques pour
l'extraction de motifs intéressants dans les données
Université Clermont 2 / LIMOS
Accueil à l'INSA Lyon / LIRIS, équipe BD (2006)

2006-2007

ATER en informatique
INSA Lyon / LIRIS

2003

Diplôme d'ingénieur en informatique ISIMA
DEA Informatique, Productique et Imagerie Médicale

- Enseignement en informatique :

- 4 enseignants-chercheurs
- Licence mathématiques-informatique
- DEUST Génie Informatique et Electronique des Systèmes (GIES)

- Recherche :

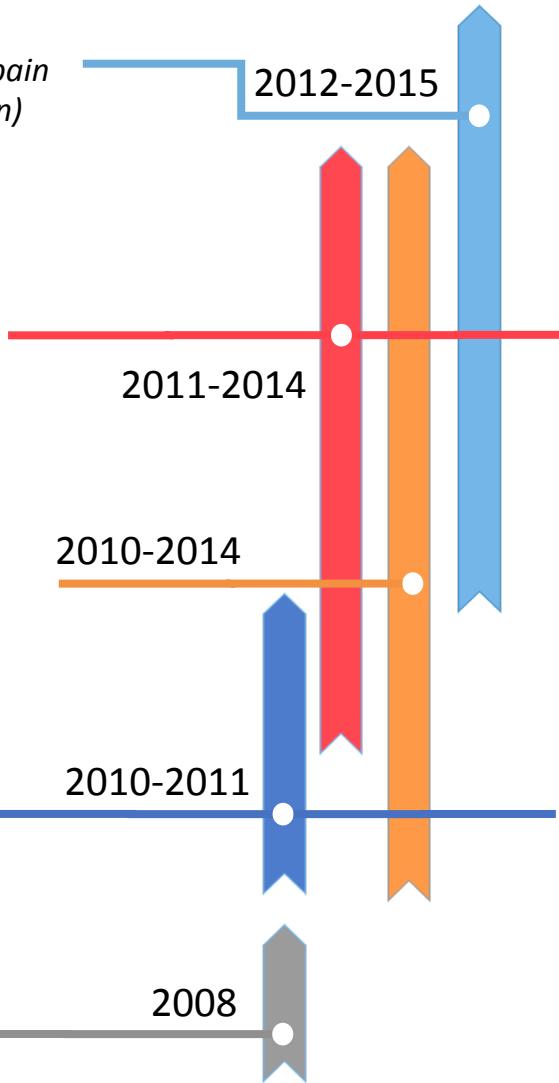
- Laboratoire renouvelé à 50%
- 13 enseignants-chercheurs (physique, chimie, géologie et informatique)
 - Forte pluridisciplinarité
- Equipe informatique
 - Une seule personne faisant activement de la recherche
 - Axe "extraction de connaissances" créé en 2006 par N. Selmaoui-Folcher
- Intégration dans des projets et participation à des encadrements de stages/thèses



Parcours recherche (1/2)

Projet CNRS Mastodons Amadouer

Analyse de données de l'environnement et l'urbain
Porteur: S. Servigne et J.-F. Boulicaut (INSA Lyon)



Jérémy Sanhes (2011-2014) :
Contribution à la fouille de données spatio-temporelles: application à l'étude de l'érosion.
Taux de co-encadrement: 40%
Direction: N. Selmaoui-Folcher (UNC) et J.-F. Boulicaut (INSA Lyon)

ANR FOSTER

Fouille de données spatio-temporelles:
application à l'érosion des sols
Porteur: N. Selmaoui-Folcher (UNC)

Projet CNRT FPBV

Fonctionnement des petits bassins versants
Porteur: M. Allenbach (UNC)

Projet Ministère Outre-Mer Dengue

Dynamique de la Dengue dans Nouméa
Porteur: M. Mangeas (IRD)

Projet industriel EDF R&D
Traitement de flux d'événements
Porteur: J.-M. Petit (INSA Lyon)

Hugo Alatrista Salas (2009-2012) :
Extraction de relations spatio-temporelles à partir des données environnementales et de la santé.
Taux de co-encadrement: 20%
Direction: M. Teisseire (IRSTEA) et N. Selmaoui-Folcher (UNC)

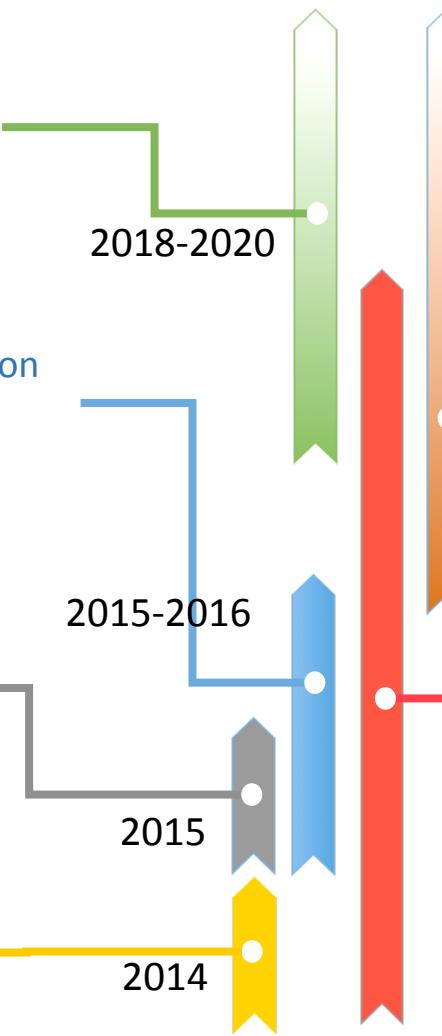
Parcours recherche (2/2)

Projet Fonds Pacifique – PIURN SOSPADIS

Dynamique des habitats informels ("squats") à Fidji et au Vanuatu

Porteur: T. Gaillard (*Ecosophy*)

Porteur PIURN: F. Flouvat (UNC)



Projet Labex Corail - DigitalGlobe Foundation

Suivi d'un récif corallien par imagerie satellitaire

Porteur: F. Flouvat (UNC)

Projet IAC Roussettes

Suivi des trajectoires des roussettes

Porteur: N. Selmaoui-Folcher (UNC)

Projet industriel KNS

Entrepôt de données et SOLAP

Porteur: N. Selmaoui-Folcher (UNC)

Jannaï Tokotoko (2017-) : Big data et science des données pour le suivi de l'aquaculture.

Taux de co-encadrement: 15%

Direction: N. Selmaoui-Folcher (UNC)

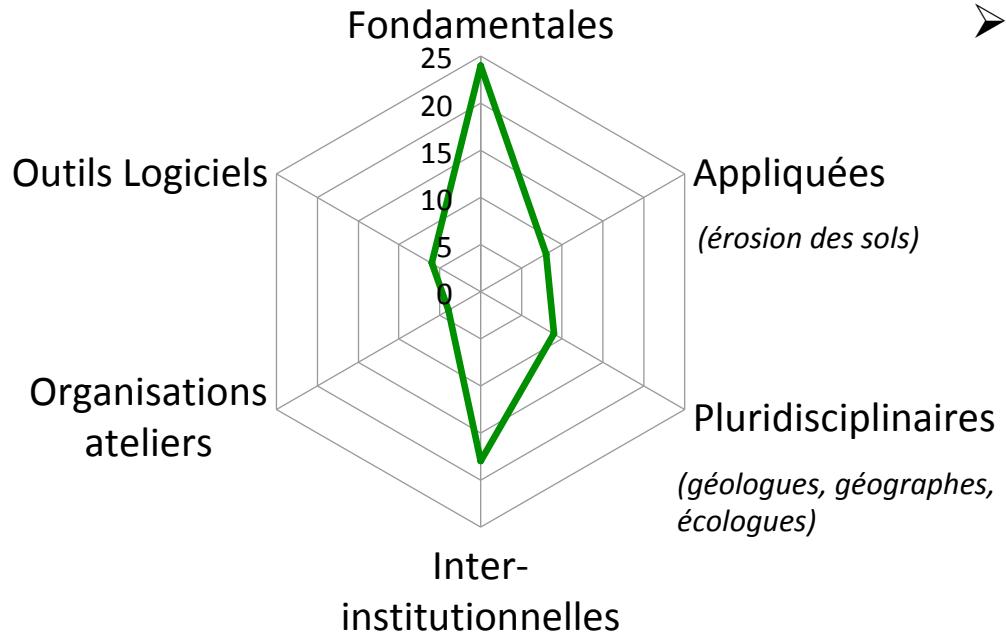
Zhi Cheng (2014-2018) : Mining recurrent patterns in a dynamic attributed graph : application to aquaculture pond monitoring by satellite images.

Taux de co-encadrement: 30%

Direction: N. Selmaoui-Folcher (UNC)

- En 2013, 2014 et 2019 : **Responsable pédagogique** de la Licence Informatique et membre du conseil de département
- Depuis 2017 : **Responsable des crédits pédagogiques** de mathématiques et informatique
- Depuis 2009 : Responsable du tutorat en informatique
- Co-porteur en 2018 d'un **projet à la fondation de l'université** intitulé « **Développement de l'esprit FacLab dans la licence informatique** » avec T. Quiniou (ISEA)

- Depuis 2017 : **Référent de l'axe "Complexité et science des données"** de l'ISEA
- Porteur du **projet PIURN SOSPADIS** (2018-2019) avec l'USP (Fidji) et Ecosophy
- Porteur du **projet Labex - DigitalGlobe Foundation** "Suivi des récifs coralliens par imagerie satellitaire" (2015-2016) avec l'EPHE et le CRIODE
- **Responsable de la tâche 1/5 "Préparation des données et validation par les experts"** du projet ANR FOSTER (2011-2014)



(INSA Lyon, Université de Montpellier, IRSTEA,
CNRS, EPHE, Université de la Polynésie Française)

- Co-encadrements :
 - 4 Thèses
 - 2 Ingénieurs CDD
 - 12 Stages niveau Master

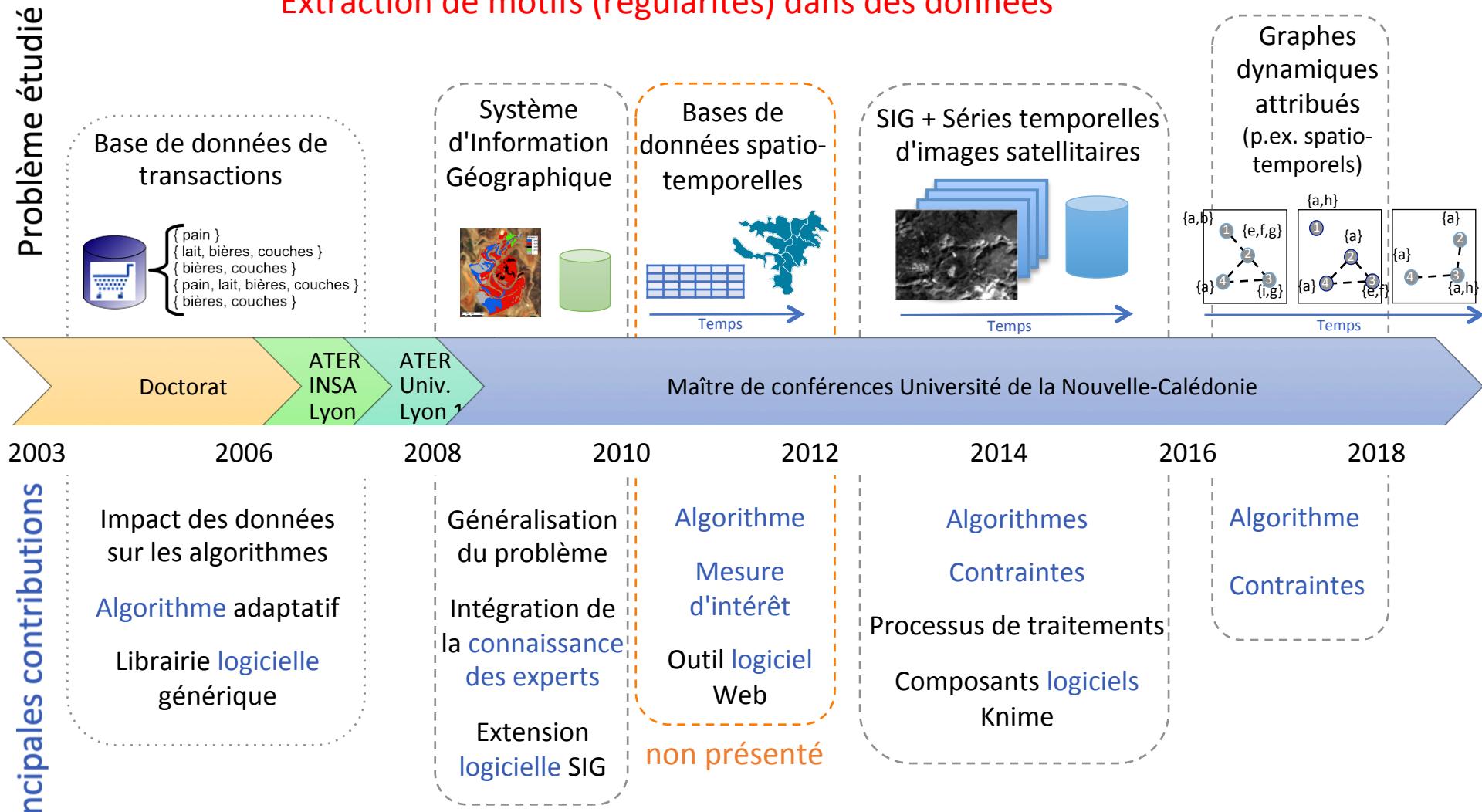
➤ 44 publications avec comité de lecture :

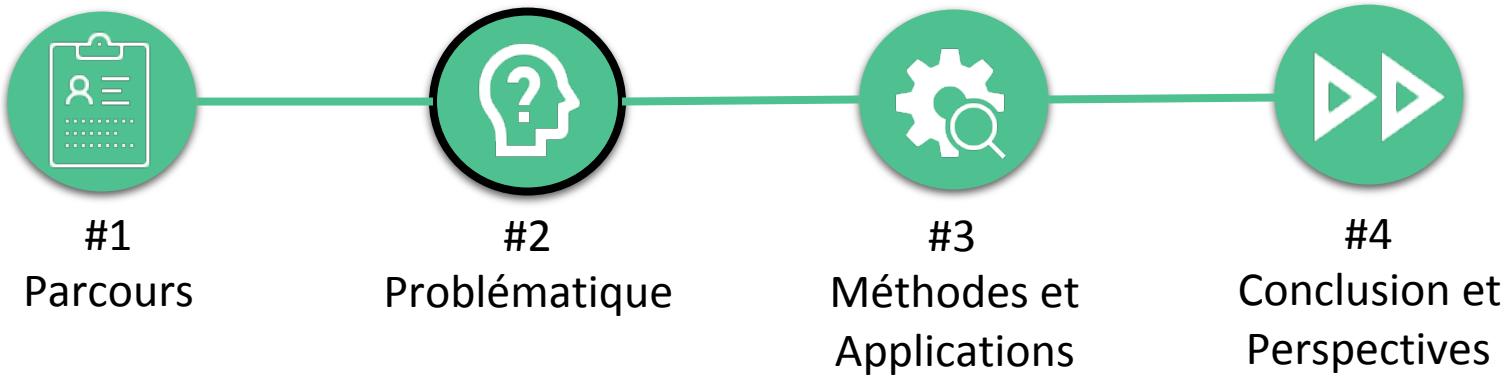
Articles dans des revues internationales ou nationales avec comité de lecture (ACL)	13
Communications avec actes dans un congrès international (C-ACTI)	16
Communications avec actes dans un congrès national (C-ACTN)	15
Direction d'ouvrages ou de revues (DO)	1
<i>Conférences données sur invitation dans un congrès national ou international (C-INV)</i>	6
<i>Diffusions de la culture scientifique</i>	9

- Membre du comité de programme de la conférence nationale EGC depuis 2010
- Relecteur régulier pour les revues internationales KAIS et KBS depuis 2011

Positionnement scientifique

Extraction de motifs (régularités) dans des données





Introduction à l'extraction de motifs

- Exemple de l'analyse du panier d'achat dans un supermarché

Transactions (ou objet)	Articles (ou propriétés)
#1	pain
#2	lait, bières, couche
#3	bières, couches
#4	pain, lait, bières, couches
#5	bières, couches

Quels sont les articles fréquemment achetés ensemble?

- Réorganisation des rayons
- Fidélisation de la clientèle
- Amélioration de la gestion du stock
- Recommandations (p.ex. Amazon)
- ...



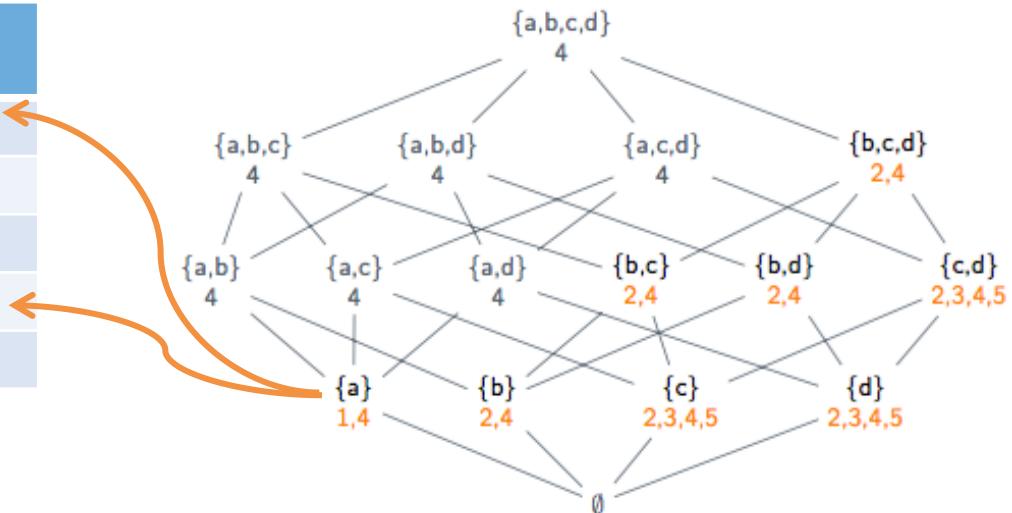
- Extraction d'ensembles d'articles fréquents [Agrawal et Srikant, VLDB'94]
 - **Données** : base de données transactionnelles
 - **Motifs** : ensembles d'articles (*itemsets*)
 - **Contrainte (ou prédictat)** : « être fréquent » (i.e. nombre de transactions avec ce motif \geq seuil)
- **Comment trouver tous ces motifs fréquents ?**
- Approche naïve : énumérer toutes les combinaisons et compter
 - {pain, lait}, {pain, bières}, {pain, couches}, ..., {lait, bières}, {lait, couches}...
 - {pain, lait, bières}, {pain, bières, couches}, {lait, bières, couches} ...
 - ...

➤ Espace de recherche exponentiel en la taille de l'entrée

- Nombre de solutions dans le pire cas : 2^n
 - p.ex. n=20 articles -> $2^{20} = 1\ 048\ 576$ solutions potentielles

Transaction (ou objet)	Articles (ou propriétés)
#1	a
#2	b, c, d
#3	c, d
#4	a, b, c, d
#5	c, d

fréquence ≥ 2



➤ Nécessité de développer des algorithmes performants

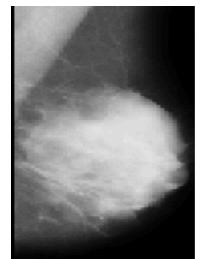
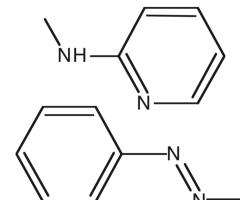
- Des **structures de données optimisées**
- Des **stratégies d'exploration efficaces**
- Des **contraintes avec des propriétés théoriques** permettant d'élaguer l'espace de recherche
 - p.ex. monotonie de la fréquence (décroissance)
 - ➔ « tout sur-ensemble d'un *itemset* non-fréquent est non-fréquent »

- Education :

- Recherche de motifs et de règles mettant en avant les différences entre des groupes d'étudiants [Minaei-Bidgoli et al., ICMLA'04]
 - "les étudiants ayant des notes moyennes au lycée, et qui font leurs devoirs maison, ont 83% de chances de réussir à l'université"
- Recherche de motifs expliquant les bons résultats de certains étudiants en mathématiques (Finlande) [Saarela et al., EDM'14]
 - "les étudiantes ont moins confiance en leurs compétences en mathématiques et ont moins l'intention de continuer dans ce domaine, bien qu'elles aient de bons résultats"

- Santé :

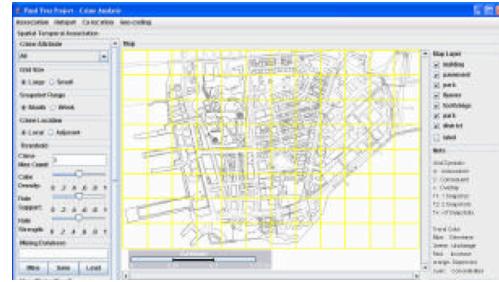
- Détection d'anomalies dans des mammographies [Antonie et al., MDM/KDD'01]
- Prédiction de la toxicité de certaines molécules [Blinova et al., Bioinformatics'03]
- Recherche de motifs rares liés à des maladies cardio-vasculaires [Szathmary et al., Int. J. Softw. Inform.'10]
 - "les sujets présentant un certain allèle ont plus de risques d'avoir un syndrome métabolique" (problèmes cardiovasculaires et diabète)



Autres applications des *itemsets* (2/2)

- Sécurité / Justice :

- Recherche de régularités dans les crimes d'un quartier de Hong-Kong [Ng et al., ADC'07]
- Découverte de discriminations pour l'obtention de crédits [Ruggieri et al., TKDD'10]
 - "discrimination des personnes âgées habitant dans un quartier avec une criminalité élevée, n'ayant pas de carte de crédit et un emprunt de 30 à 42 mois"



- Développement logiciel :

- Découverte d'erreurs de programmation communes [Livshits et al., FSE'05]
- Analyse des pratiques lors du développement d'un logiciel [Sun et al., ICSSP'13]

- Cyber-sécurité / Détection de fraudes :

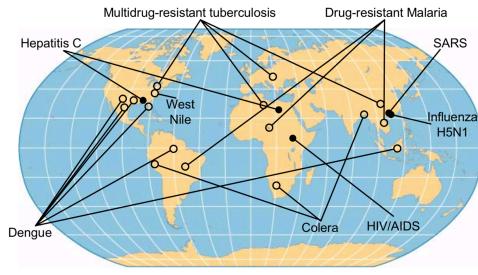
- Détection d'achats frauduleux par des cartes bancaires [Sanchez et al., ESWA'08]
 - "les jeunes hommes sont plus affectés par les fraudes"
- Détection de virus informatiques dans des téléphones mobiles [Coletta et al., FC'16]
 - Détection d'une attaque ciblant les transactions bancaires de clients Coréens et Chinois

Applications liées à des données spatiales

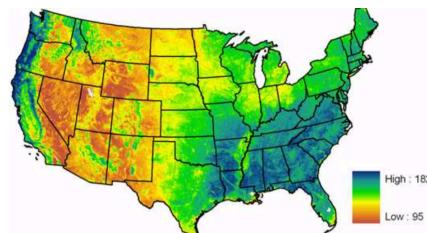
- Explosion de la quantité et de la diversité des **données spatiales** collectées avec de plus en plus de **suivis dans le temps**



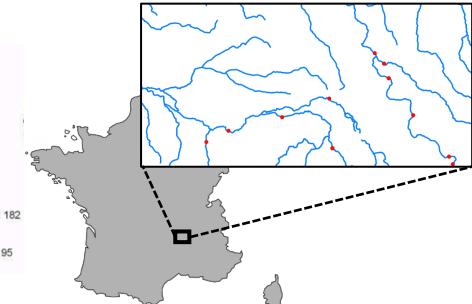
Données de mobilités
(trajectoires)¹



Données d'événements²



Informations continues
sur des régions³



Données de réseaux

- Un grand nombre d'applications différentes

- Etude des mouvements migratoires d'animaux [Phan et al., IDA'12]
- Analyse du déplacement des touristes [Zhou et Meng, DASFAA'11]
- Suivi d'ouragans [Lee et al., SIGMOD'07]
- Etude de la propagation de maladies [Alatrista et al., PAKDD'12]
- Prévention de la criminalité dans une ville [Celik, KAIS'15]
- Analyse du trafic automobile [Liu et al., KDD'11]
- Suivi de l'activité du soleil [Aydin et Angryk, ICDMW'16]
- ...

1. [Phan, PhD thesis 2013]

2. [Fauci, R.H. Erbert Memorial lecture 2006]

3. [Ding et al., SDM'09]

Extraction de motifs mais généralement pas des *itemsets*
(pas assez riches)

Evolution des travaux existants

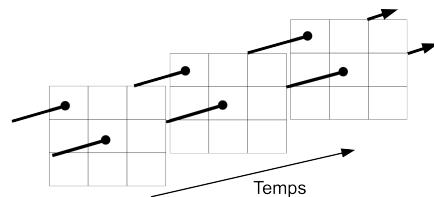
- Vers des motifs de plus en plus riches dans des données toujours plus importantes et complexes

Ensemble de prédicts spatiaux
 [Kopersky et al., SSD'95]

$\text{is_a}(X, \text{city}) \wedge \text{within}(X, BC) \wedge \text{adjacent_to}(X, \text{water})$

Evolution de zones (séquences)
 [Tsoukatos et al., SSTD'01]

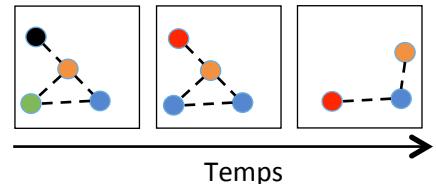
`< {pluie_forte, pente}, {érosion} >`



Graphes de voisinage dans une image satellitaire
 [Ozdemir et al., ICPR'10]



Graphes planaires spatio-temporels dans une série d'images
 [Prado et al., IDA'13]



1993

1998

2003

2008

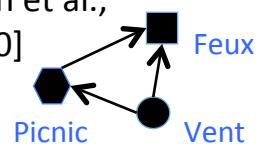
2013

2018

Co-localisations
 [Shekhar et Huang, SSTD'01]



Graphes orientés acycliques (DAG) d'évènements
 [Mohan et al., SDM'10]



Problème

Un grand nombre d'avancées mais encore des difficultés pour intégrer, analyser et croiser toutes les données disponibles

- spatiales **ET** temporelles **ET toutes** les informations sur les objets étudiés

Verrous scientifiques et contributions

Suivi environnemental

- des phénomènes complexes (p.ex. érosion)
- une connaissance des experts (empirique ou formelle)
- besoins d'outils et de solutions facilement interprétables

Contexte applicatif



Données



Verrous

Extraction de motifs spatio-temporels plus riches et plus pertinents



Contraintes

- définies par les experts
- dérivées de modèles

- numériques
- nominales
- spatiales
- temporelles

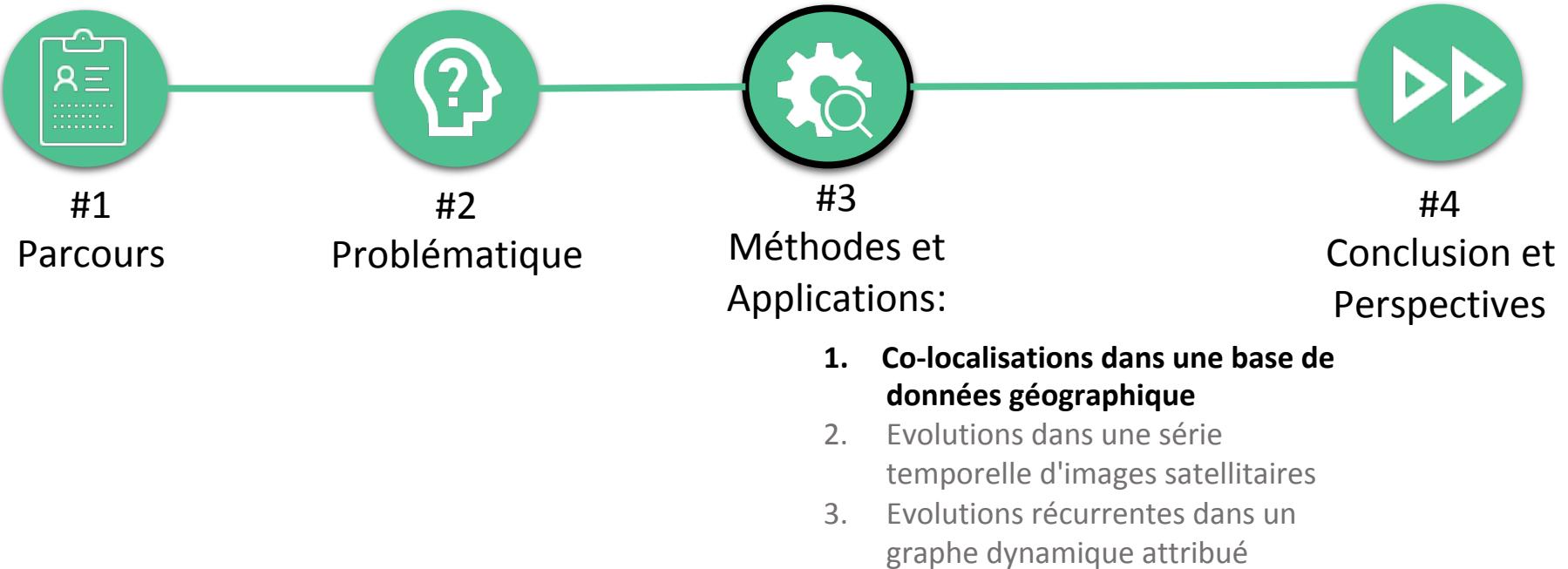
- bases de données
- SIG
- images

Outils pour extraire et visualiser des

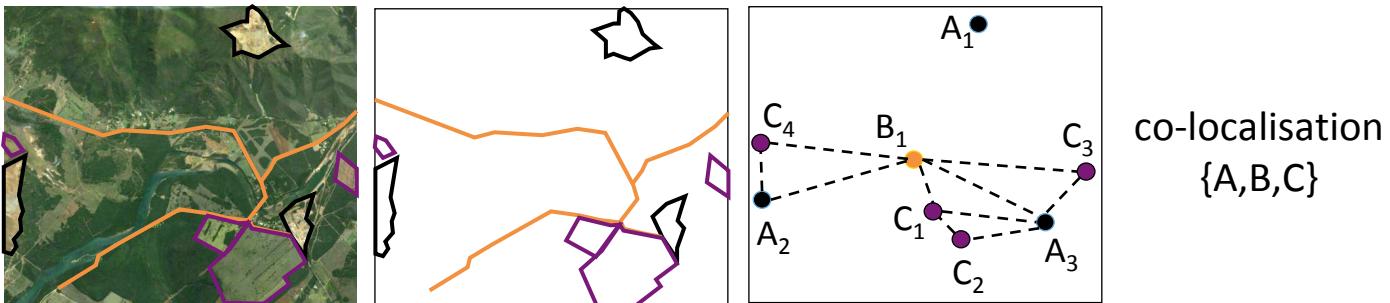
- co-localisations dans un SIG
- évolutions de quartiers
- évolutions dans une série d'images satellitaires

Algorithmes d'extraction de

- motifs spatio-séquentiels
- chemins dans des DAG attribués
- graphes dynamiques attribués



- **Problème :** Extraire des ensembles d'évènements/objets souvent localisés à proximité
 - **Données :** base de données géographiques



- **Motifs :** ensembles d'objets/événements
- **Contrainte :** « être souvent proche »
- De nombreux travaux par rapport à ce problème, p.ex. [Huang et al., TKDE'04] [Bogorny et al., ICDM'06] [Celik et al., TKDE'08]
 - Optimisation de l'extraction (algorithme et structure de données)
 - Proposition de nouvelles contraintes/mesures (spatiales, temporelles, et statistiques)
- **Limite des approches existantes :**
 - Pas ou peu d'intégration de la connaissance du domaine (post-traitements)
 - Pas de méthode de visualisation graphique et synthétique des motifs extraits

- Proposition : Intégrer des contraintes permettant d'exclure des co-localisations connues ou non-intéressantes pour les experts
 - Des solutions moins nombreuses et plus pertinentes
 - Une intégration dans les algorithmes d'extraction pour en améliorer les performances
- Utilisation du cadre théorique de [Mannila et Toivonen, 97] pour faire de l'**extraction de co-localisations sous-contraintes des experts**
 - Possibilité d'utiliser tout une famille d'algorithmes d'extraction
- Proposition de contraintes **spatiales** et **thématiques** identifiées par les experts et **vérifiant la propriété de monotonie**

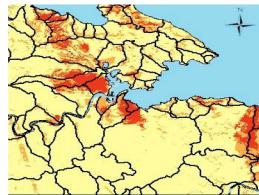
Contraintes thématiques	Type
$\neg(F \subseteq \varphi)$	objets/événements
$\neg(\varphi \cap F \neq \emptyset)$	objets/événements
$\neg(\varphi \cap t \geq 2)$	intra-thèmes
$\neg(\forall t \in T(\varphi \cap t \neq \emptyset))$	thème et inter-thèmes
$\neg(\exists t_1 \in T(\varphi \cap t_1 \neq \emptyset) \wedge \exists t_2 \in T(\varphi \cap t_2 \neq \emptyset))$	inter-thèmes partiel

Contraintes spatiales	Type
$(\forall o \in I(r(o, shape)))$	toute relation spatiale booléenne
$(\forall o \in I(r(o, shape))) \vee CF$	spatiale et thématique
$(\forall o \in I(r(o, shape))) \wedge CF$	spatiale et thématique

avec F un ensemble d'objets, t un thème, T un ensemble de thèmes, I une instance de la **co-localisation φ à tester**

- Possibilité de combiner toutes ces contraintes (conjonction de contraintes)
 - p.ex. extraire les co-localisations fréquentes associées aux thèmes de l'érosion et des constructions humaines dans certaines zones géographiques

- **Limite de contraintes définies par des experts** : une définition "manuelle" des contraintes nécessitant un investissement important des experts



- **Idée** : Exploiter la connaissance "encodée" dans les modèles définis dans la littérature du domaine et en dériver des contraintes
- Beaucoup de modèles exprimés sous la forme de fonctions de plusieurs variables

Exemple pour l'érosion des sols

Modélisation du détachement par goutte de pluie dans le modèle d'érosion RMMF

$$F(x_K, x_R, x_A, x_{CC}, x_I, x_{PH}) = x_K[x_R \cdot x_A(1-x_{CC})(11.9 + 8.7 \log x_I) + (15.8 + x_{PH}^{0.5}) - 5.87] \cdot 10^{-3}$$

x_K indice de détachement du sol
 x_R précipitation annuelle
 x_I intensité de la pluie

x_A proportion de pluie interceptée par la canopée
 x_{CC} pourcentage de couverture de la canopée
 x_{PH} hauteur de la végétation

Exemple de contrainte liée à l'érosion : $F(\varphi) \geq \text{minf}$, i.e. se focaliser sur les motifs liés à une forte perte en sol (risque d'érosion)

- Des motifs validés par les données et le modèle

Intérêts : ➤ Des contradictions entre les données et les sorties du modèle

- Des corrélations avec d'autres variables non prises en compte dans le modèle mais présentes dans les données

- Difficulté 1 : Exploiter le modèle F si certaines variables ne sont pas présentes dans les données ou si certaines valeurs sont approximatives (i.e. des intervalles)

co-localisation $\varphi = \{"x_K=\text{latérite}", "x_R=6000", "x_A=0.3", "x_{CC}=0.1", "x_I=25"\}$

➤ $F(\varphi) = F(4, 6000, 0.3, 0.1, 25, ?) = ?$

➤ Quelle valeur mettre pour x_{PH} (hauteur de la végétation) ?

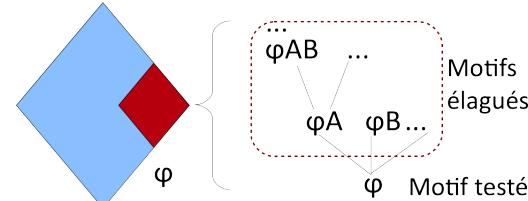
- Calculer une borne inférieure et supérieure pour $F(\varphi)$

$F(4, 6000, 0.3, 0.1, 25, 0) \leq F(\varphi) \leq F(4, 6000, 0.3, 0.1, 25, 130)$, car $x_{PH} \in [0, 130]$ et F croissant

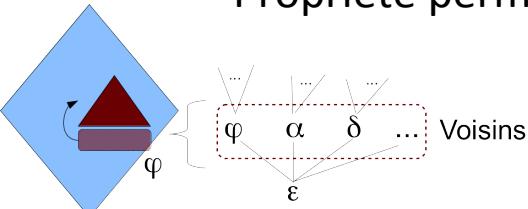
- Difficulté 2 : Identifier des propriétés théoriques permettant d'éviter de tester certains motifs lors de la recherche (optimisation des algorithmes)

- Propriété permettant d'élaguer des sur-ensembles

• si $F(\varphi) < \min_f$ alors tout sur-ensemble aussi



- Propriété permettant d'élaguer des motifs "voisins"



Soient $\varphi = \{\dots, "x_{PH}=[5,10["\}$ et $\alpha = \{\dots, "x_{PH}=[1,5]" \}$ avec "..." identiques et F croissant par rapport à x_{PH} sur $[1, 10]$
si $F(\varphi) < \min_f$ alors $F(\alpha)$ aussi

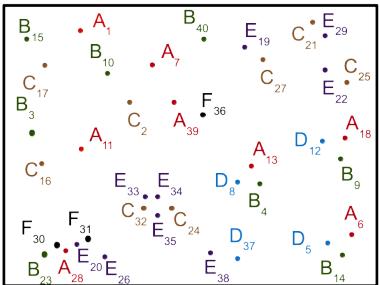
Une visualisation intuitive des co-localisations

➤ Proposition : visualiser les co-localisations extraites dans un SIG (en post-traitement)

- Une visualisation cartographique intuitive pour les experts
- Des informations spatiales et thématiques supplémentaires

co-localisations
{E,C}
{A, B, C}
{A, B, D}
etc.

?



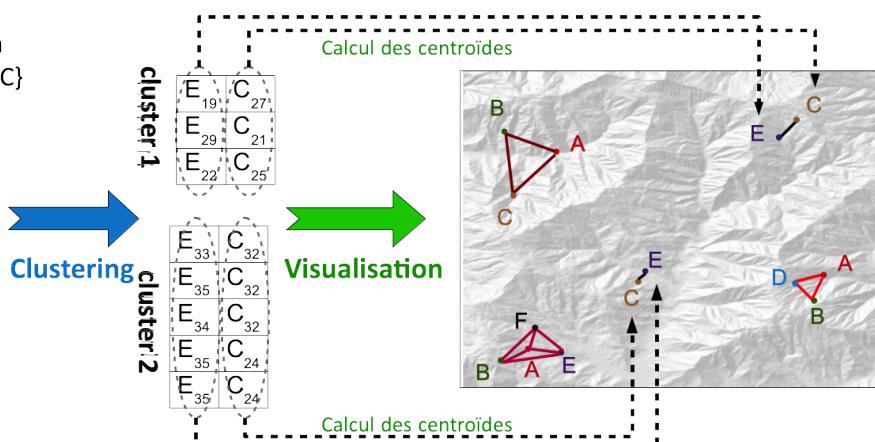
- Comment les représenter graphiquement ?
- Quelle localisation leur affecter ?

➤ Faire un *clustering* spatial des occurrences de chaque co-localisation et utiliser les centroïdes pour les représenter

- Montre où et comment chaque co-localisation est spatialement distribuée

Occurrences de la
co-localisation {E,C}

E ₁₉	C ₂₇
E ₂₉	C ₂₁
E ₂₂	C ₂₅
E ₃₃	C ₃₂
E ₃₅	C ₃₂
E ₃₄	C ₃₂
E ₃₅	C ₂₄
E ₃₅	C ₂₄



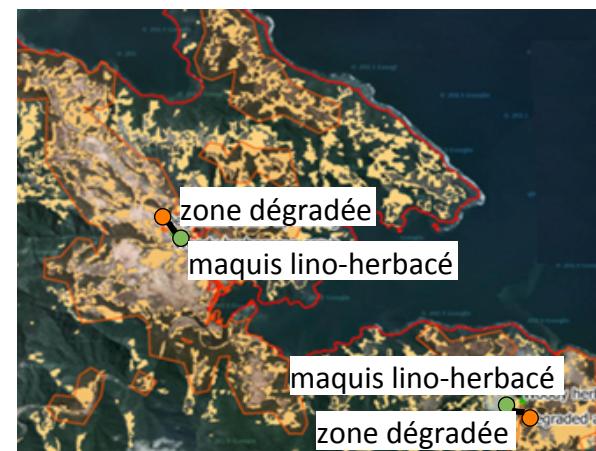
Extraction de co-localisations guidée par le domaine : Quelques résultats (1/2)

- Développement d'un prototype intégrant contraintes du domaine et visualisation des solutions dans PostGIS
- Expérimentations sur des données liées à l'érosion des sols en Nouvelle-Calédonie

données	nb objets spatiaux	nb de types d'objets
"Ouinné"	3 943	68
"Kwe Binyi"	7 306	71

➤ Exemples de motifs extraits :

- {"zone dégradée", "maquis lino-herbacé"} : association intéressante en raison du fort taux d'endémisme (50% des cas)
- {"latérites épaisses", "maquis lino-herbacé", "pente=[3.6;30]"} : motif apparaissant dans 5% de la région et associé à un risque d'érosion par un modèle du domaine

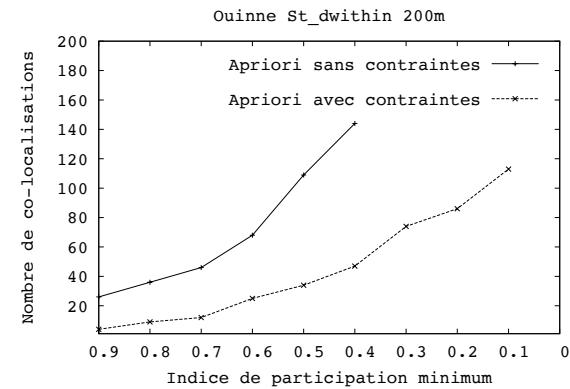
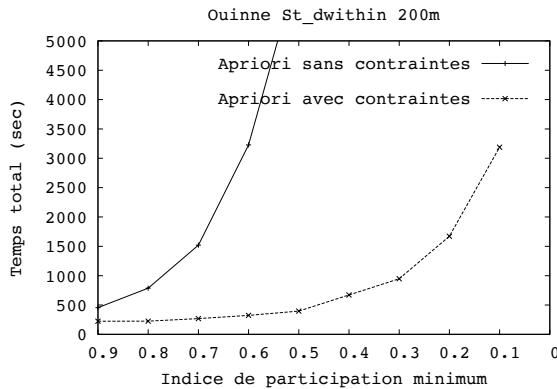


Extraction de co-localisations guidée par le domaine : Quelques résultats (2/2)

➤ Impact des méthodes proposées sur les performances et le nombre de solutions affichées

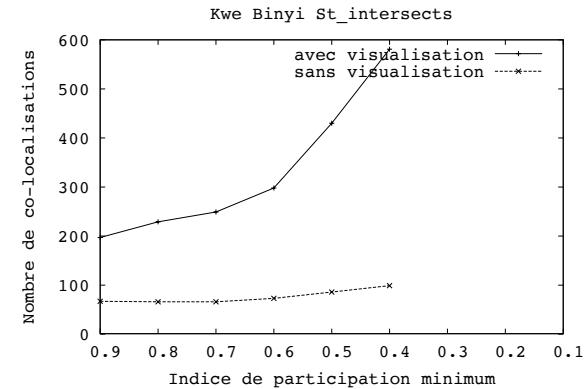
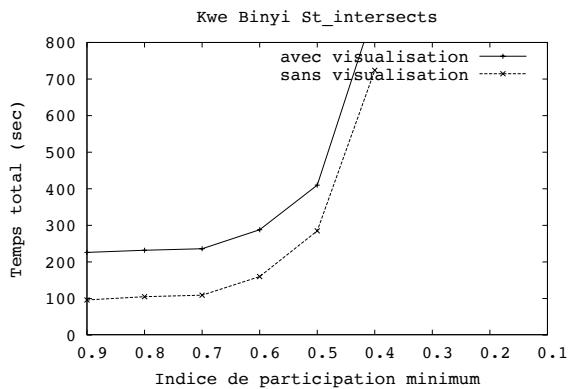
- Contraintes du domaine

Forte diminution du temps d'exécution et du nombre de motifs extraits



- Visualisation

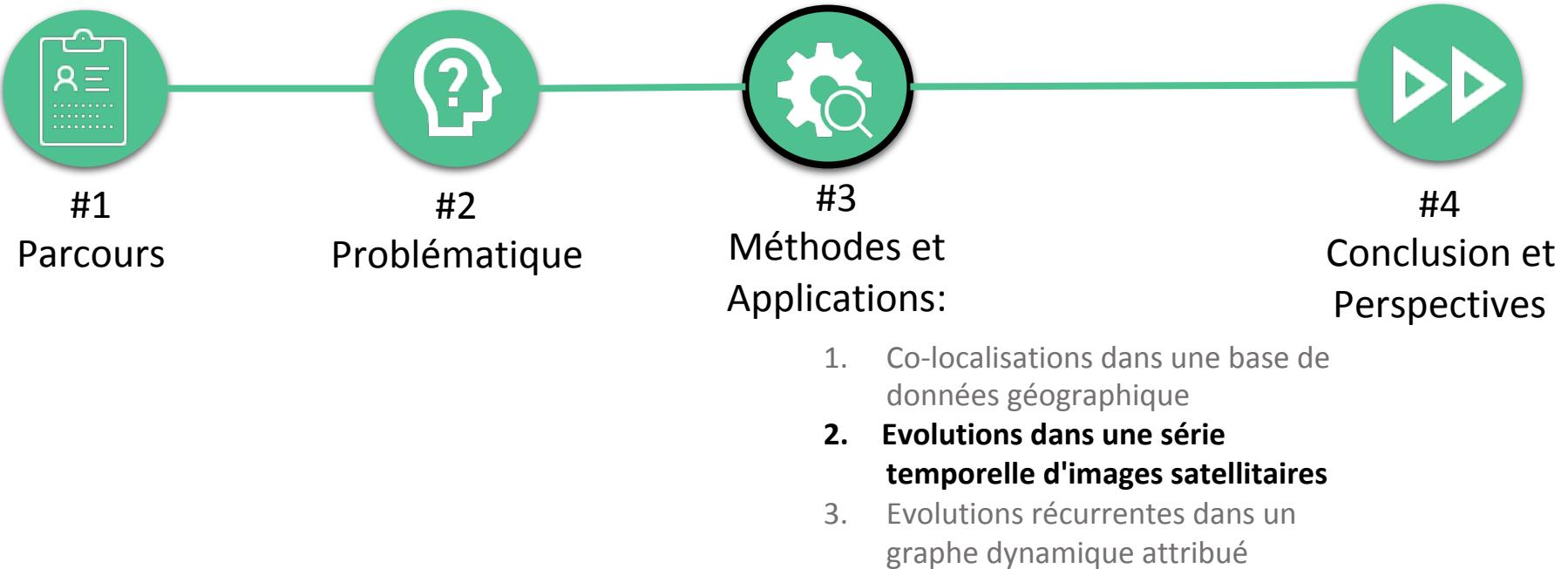
- Augmente sensiblement le temps d'exécution
- En moyenne, 3 motifs affichés par co-localisation extraite



Indice de participation minimum = "probabilité d'apparition" minimale de la co-localisation

Travaux publiés dans :

- ACM Symposium on Applied Computing (SAC), 2010
- International Journal of Agricultural and Environmental Information Systems (IJAEIS), 2011
- International Database Engineering and Applications Symposium, 2011
- European Conference on Artificial Intelligence (ECAI), 2014
- GeolInformatica, 2015

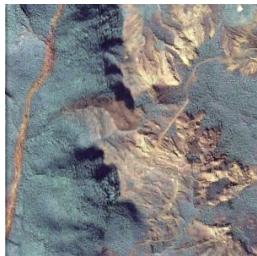


Problématique de l'extraction de motifs dans une série temporelle d'images satellitaires

- Problème: Extraire des évolutions fréquentes dans une série temporelle d'images satellitaires complétée par des données issues de SIG



Date t1

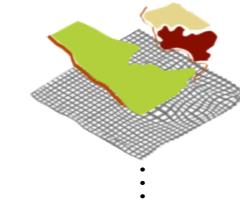
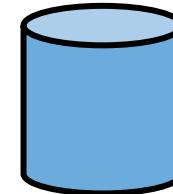


Date t2



Date t3

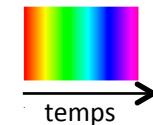
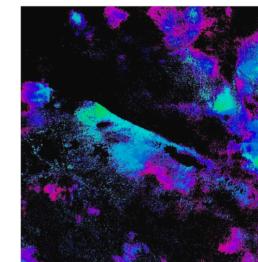
Séquence d'images satellitaires très haute résolution
(~10 Go)



Données SIG

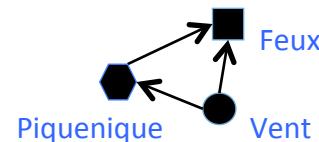
Modèle numérique de Terrain (MNT), descriptions des sols (végétation, nature géologique, etc.), météorologie, ...

- Utilisation des travaux sur les séquences spatiales,
p.ex. [Tsoukatos et al., SSTD'01], [Méger et al., PKDD'15]
 - Limite: des objets/zones géographiques figées



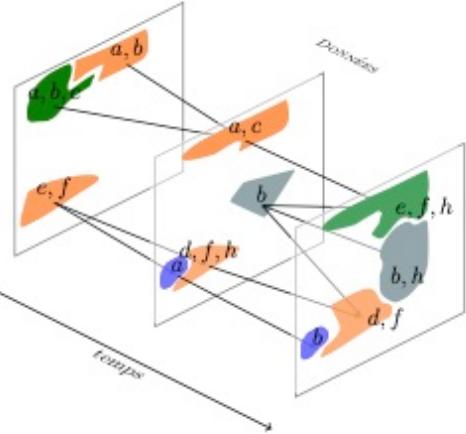
GFS-pattern A->A->B->C->A->A

- Extraction de graphes orientés acycliques (DAG) d'évènements [Mohan et al., SDM'10]
 - Limite: un seul attribut par objet



➤ Proposition :

1. Modéliser des données spatio-temporelles sous forme de DAG attribué



nœud = objet spatial décrit par un ensemble d'attributs

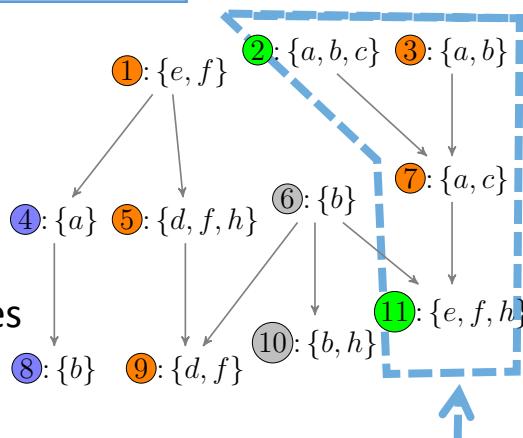
arête = relation de proximité spatio-temporelle

- Capture des **dynamiques complexes** (p.ex. apparition/disparition, fusion/division)
- Intègre **toutes les informations disponibles** (radiométriques, données SIG, etc.)
- **Générique**, i.e. d'autres applications potentielles (p.ex. réseaux sociaux, bioinformatique)

2. Extraire des chemins fréquents dans ce DAG attribué

$\{a,b\} \rightarrow \{a,c\} \rightarrow \{e,f,h\}$, ou plus simplement $ab \rightarrow ac \rightarrow efh$

- Représente l'évolution des attributs des objets dans l'espace et le temps
- Offre un **bon compromis entre expressivité et complexité d'extraction** (et d'interprétation)
 - + des séquences d'*itemsets* sous contraintes
 - des occurrences entrelacées dans un unique graphe

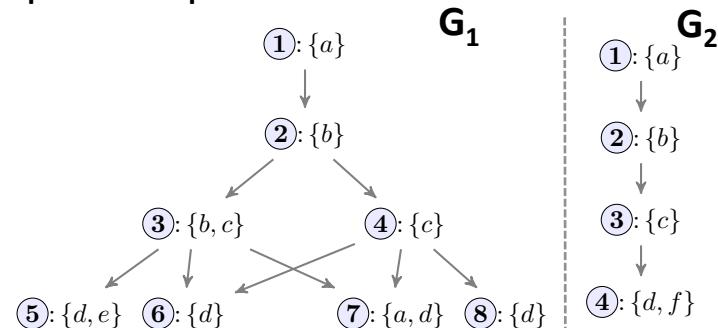


- Difficulté 1 : Définir la fréquence d'un motif dans un graphe unique

Quelle est la fréquence de $a \rightarrow b \rightarrow c \rightarrow d$ dans G_1 et dans G_2 ?

➤ Introduction de la notion de chemin pondéré

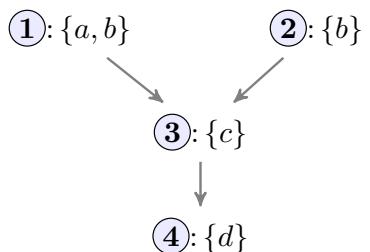
$a^{-1} \rightarrow b^{-2} \rightarrow c^{-6} \rightarrow d$ dans G_1 et $a^{-1} \rightarrow b^{-1} \rightarrow c^{-1} \rightarrow d$ dans G_2



➤ Contrainte de fréquence : poids minimum \geq seuil

- Une mesure monotone et peu coûteuse à calculer

- Difficulté 2 : Beaucoup de motifs potentiellement redondants, i.e. n'apportant pas d'information



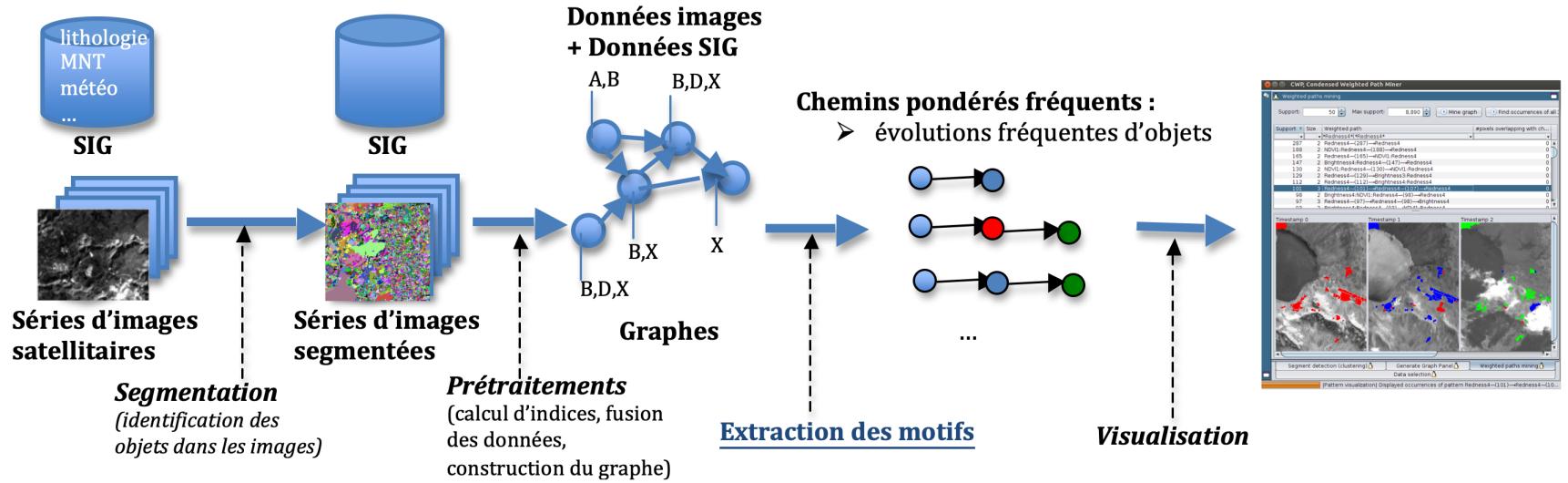
chemin pondéré	chemins pondérés plus spécifiques
$c^{-1} \rightarrow d$	$ab^{-1} \rightarrow c^{-1} \rightarrow d$ $b^{-2} \rightarrow c^{-1} \rightarrow d$

➤ Contrainte de non-redondance : ne conserver que les motifs les plus spécifiques

- Représentent les mêmes données (l'évolution des mêmes objets)
- Possibilité de retrouver tous les autres à partir de ceux-ci (sans accès aux données)

Processus d'extraction des chemins pondérés fréquents et non-redondants

➤ Mise en place d'un processus complet d'analyse de séries d'images satellites

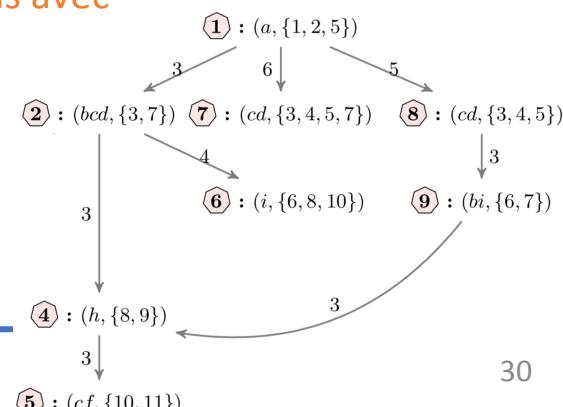


➤ Extraction des motifs suivant un parcours en profondeur, des extensions progressives des motifs et des projections successives des données

Stratégie similaire à l'algorithme *PrefixSpan* [Pei et al., ICDE'00], mais avec

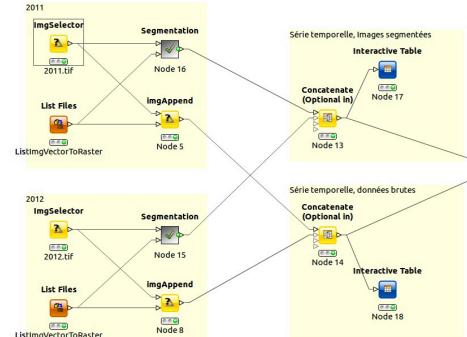
- des projections basées sur les arêtes à étendre
- un calcul d'*itemsets* fermés fréquents (localement)
- des extensions avec mise à jour du préfixe

➤ Optimisation de l'algorithme grâce à un stockage des motifs sous forme de graphe



Extraction de chemins dans un unique DAG attribué : Quelques résultats (1/2)

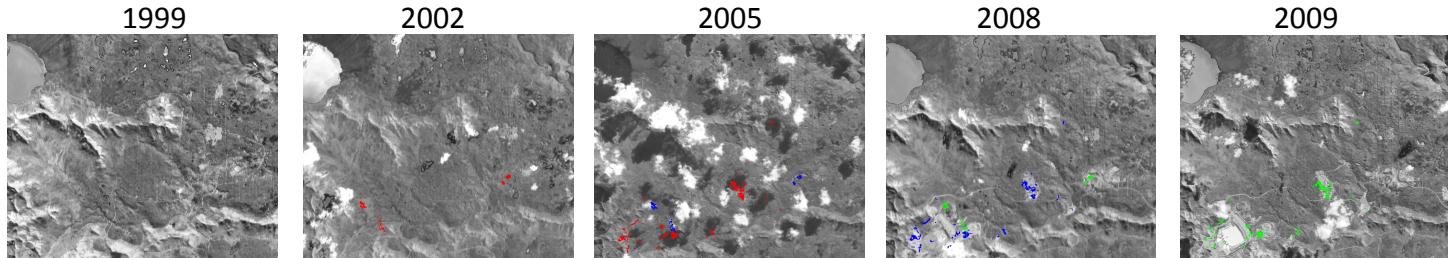
- Développement de plugins dans la plate-forme open source d'analyse KNIME
- Expérimentations sur des données synthétiques, liées à l'érosion des sols et liées à des citations de brevets



données	nb de noeuds	nb d'arêtes	nb d'attributs par noeuds	nb total de valeurs d'attributs
"érosion"	25 618	41 166	6	262
"brevets"	184 284	414 487	5-7	506
"synthétiques"	20 000 - 200 000	60 000 - 600 000	1-10	15

➤ Exemples de motifs extraits

- Redness1^{-58->}Redness4, NDVI0^{-61->}Redness4, NDVI0, Brightness4 : augmentation de l'érosion (rougeur du sol et luminosité élevée) sur 3 dates avec une végétation très faible (indice NDVI)



- NDVI2^{-57->}NDVI3^{-58->}NDVI4 : augmentation de la végétation (prolifération d'algues)

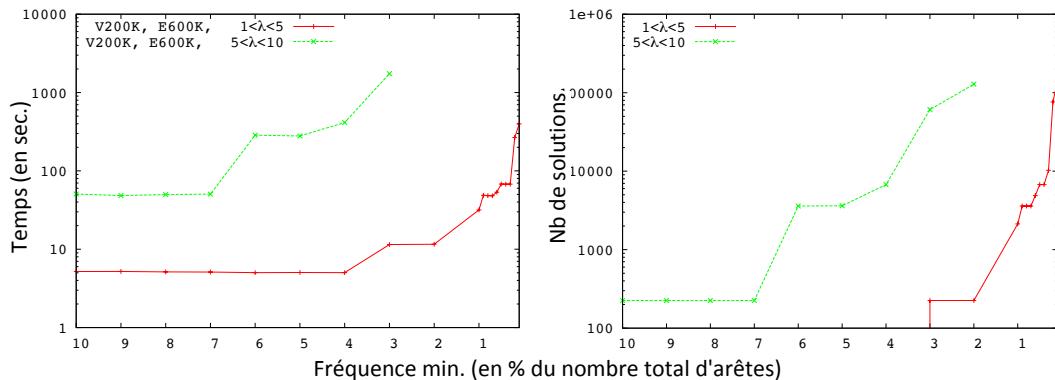
Extraction de chemins dans un unique DAG attribué : Quelques résultats (2/2)

➤ Impact des données en entrée sur les performances et le nombre de solutions

➤ Influence du nombre d'attributs par noeud

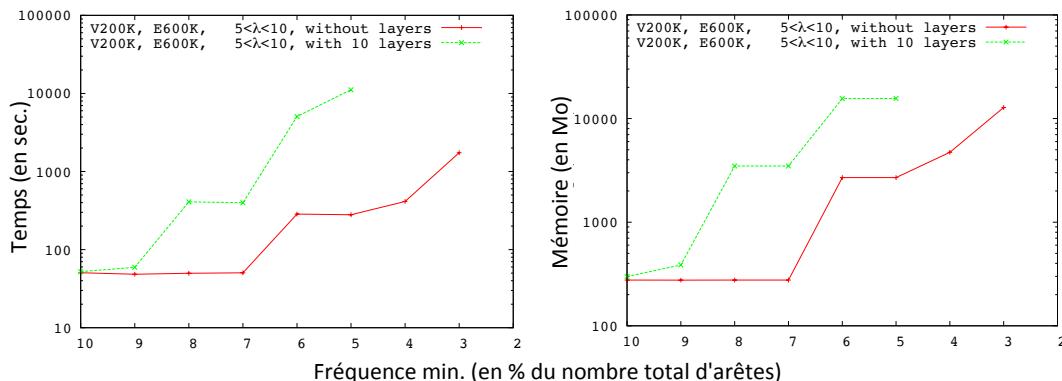
Impact fort du nombre d'attributs par noeud

➤ Illustrer la difficulté d'analyser un DAG attribué



➤ Influence de la structure du graphe (couches temporelles)

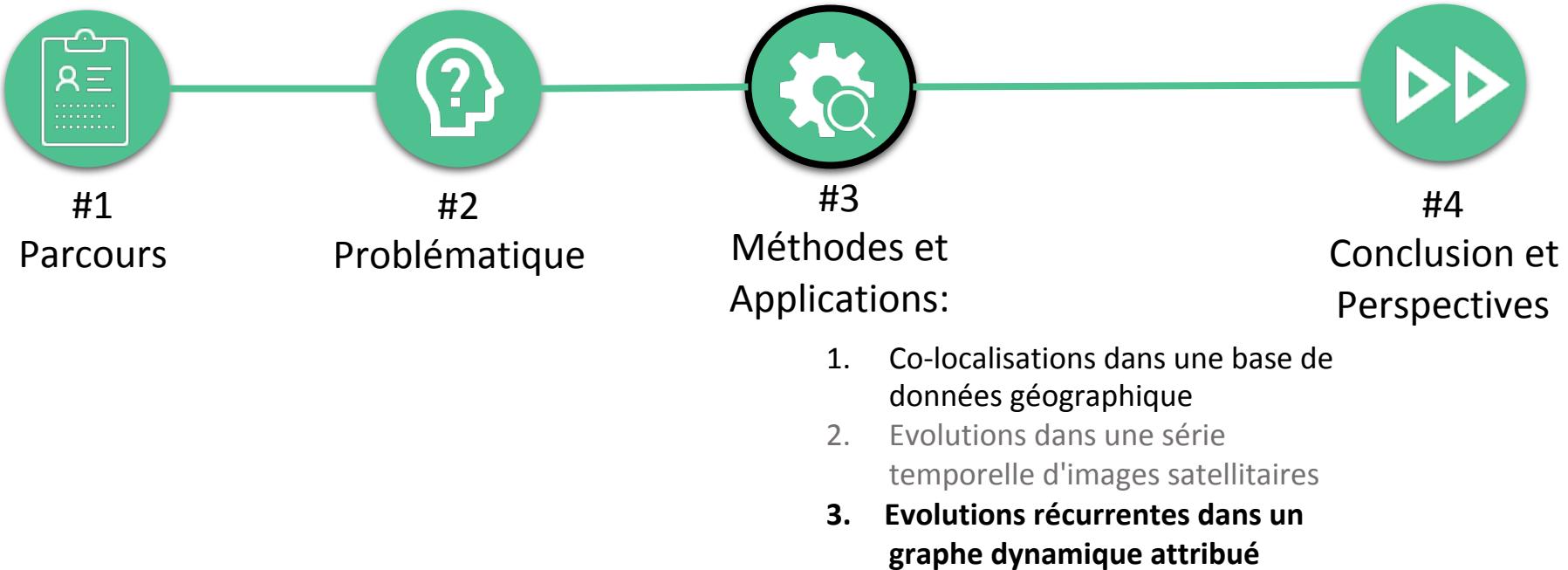
Plus difficile d'analyser un graphe issu d'une série temporelle d'images



Travaux publiés dans :

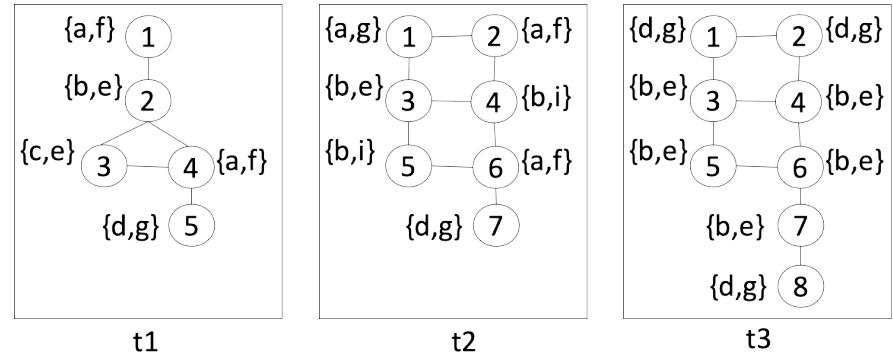
- International Conference on Database and Expert Systems Applications (DEXA), 2011
- International Joint Conference on Artificial Intelligence (IJCAI), 2013

- IEEE International Conference on Data Mining (ICDM), Demos Session, 2016



Problématique de l'extraction de motifs dans un graphe dynamique attribué

- Problème : Extraire des évolutions intéressantes dans un graphe dynamique attribué
 - Données : une séquence de graphes étiquetés par des attributs

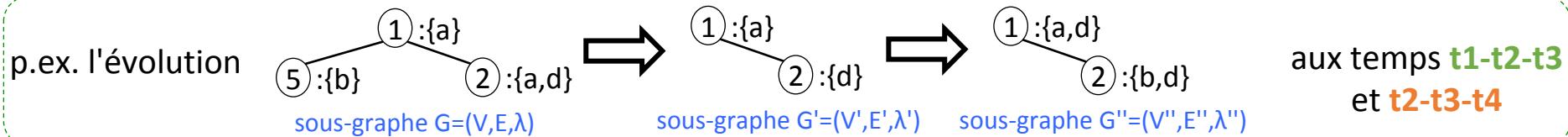
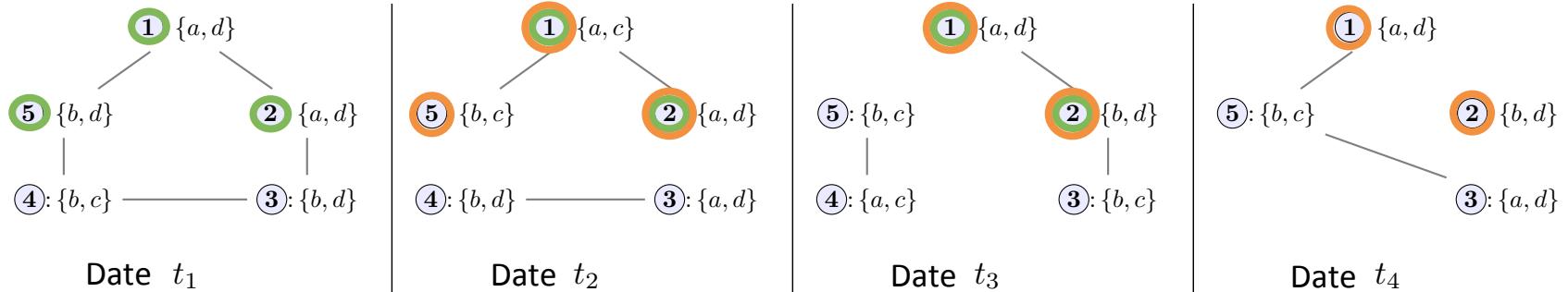


nœud = objet spatial décrit par un ensemble d'attributs

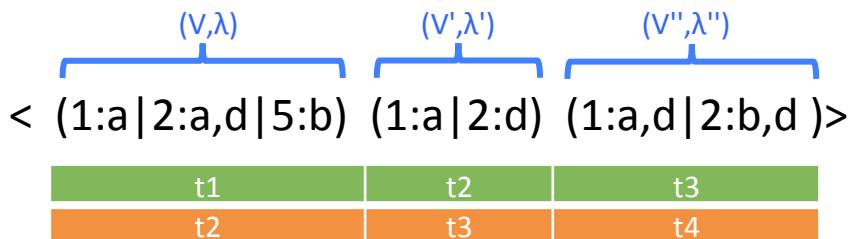
arête = relation entre deux objets (p.ex. proximité spatiale)

- Un grand nombre d'applications (p.ex. réseaux sociaux, composants chimiques, transports)
- Beaucoup de travaux sur l'extraction de sous-graphes fréquents dans une collection de graphes, p.ex. [Yan et al., ICDM'02] [Inokuchi et al., Machine Learning'03]
- Peu de travaux sur l'extraction de graphes attribués et dynamiques attribués
 - Explosion combinatoire et complexité du graphe unique
- Des domaines de motifs très spécifiques (p.ex. extraction de "communautés"), des contraintes fortes et/ou des hypothèses sur les données en entrée

➤ Proposition : Extraire des séquences de sous-graphes intéressantes dans un unique graphe dynamique attribué

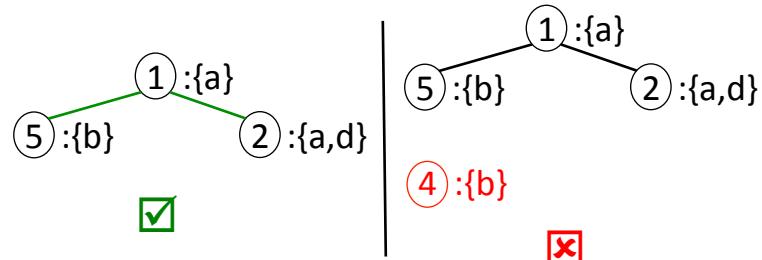


➤ Définition d'un nouveau domaine de motifs représentant l'évolution d'un sous-ensemble de noeuds, i.e. "une séquence de sous-graphes attribués sans les arêtes"

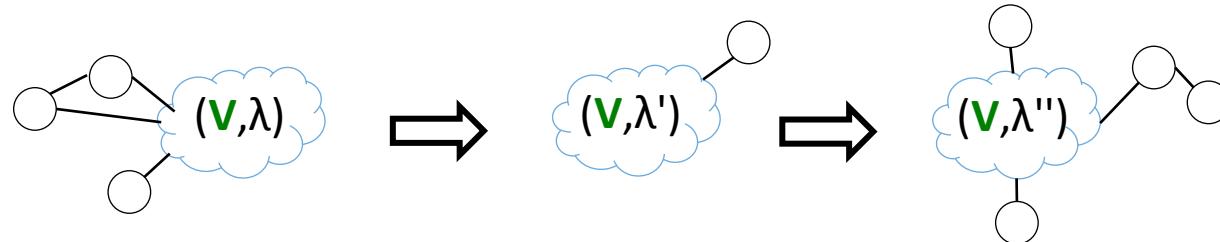


➤ Pas de prise en compte **directe** des arêtes afin d'élargir l'analyse

- Définition de plusieurs contraintes de
 - connectivité : des nœuds connexes/liés



- volume : un nombre minimum de nœuds à chaque temps
- continuité temporelle : un nombre minimum de nœuds communs à tous les temps
 - Le sous-graphe évolue autour d'un ensemble de nœuds

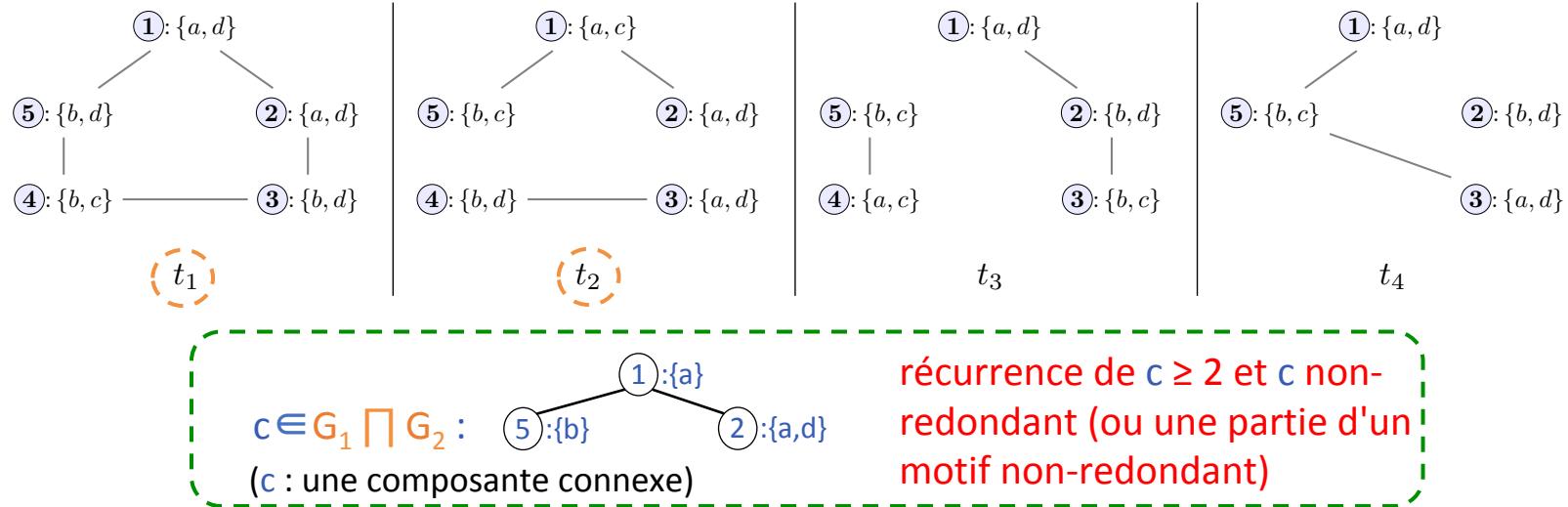


- non-redondance : élimine les solutions portant la même information (id. précédemment avec les DAG)
- récurrence : un nombre minimum de répétitions dans le temps

< (1:a|2:a,d|5:b) (1:a|2:d) (1:a,d|2:b,d)> ➤ 2 répétitions

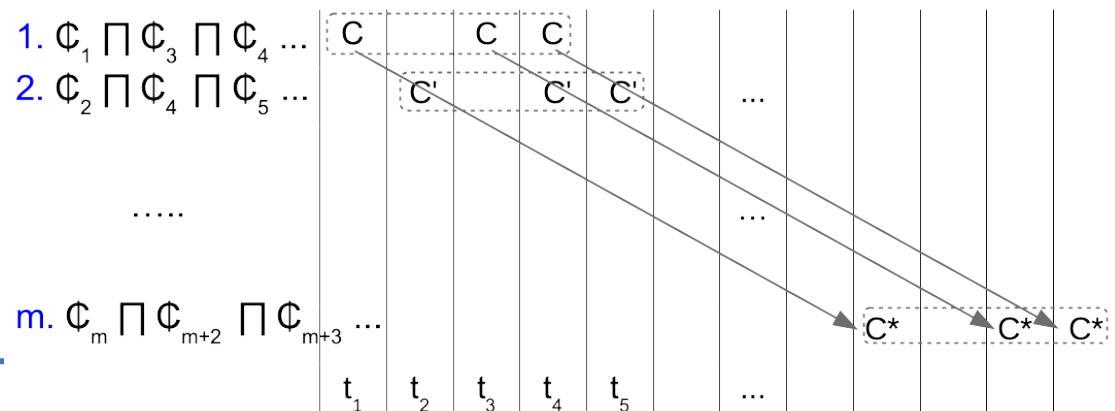
Extraction des évolutions récurrentes dans un unique graphe dynamique attribué

- Deux propriétés des contraintes de récurrences et de non-redondances par rapport aux intersections de graphes \prod (nœuds, arêtes et attributs)



- Proposition d'une **stratégie incrémentale** originale basée sur des intersections des composantes connexes contenues à chaque temps

Extensions progressives des motifs avec une génération de **motifs de tailles potentiellement différentes** à chaque extension



C_i : les composantes connexes au temps t_i

c, c' et c^* : des composantes connexes issues des intersections

Extraction de motifs dans un unique graphe dynamique attribué: Quelques résultats (1/2)

- Expérimentations sur des données synthétiques, liées au réseau social de citations DBLP et liées au trafic aérien aux USA lors de plusieurs ouragans sur la côte Est

données	nb de noeuds	nb d'arêtes	nb d'attributs par noeuds	nb de dates
"trafic aérien"	280	1 206	8	8 (semaines)
"DBLP"	2 723 (par date)	10 737 (en moyenne)	43	9 (1990 à 2009)
"synthétiques"	250 -20 000	1 000 - 80 000	50	12



➤ Exemple de motifs extraits

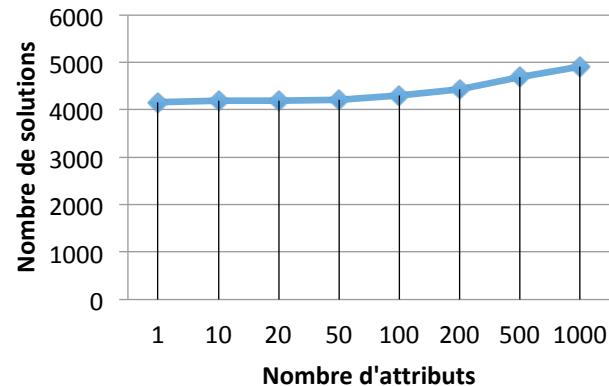
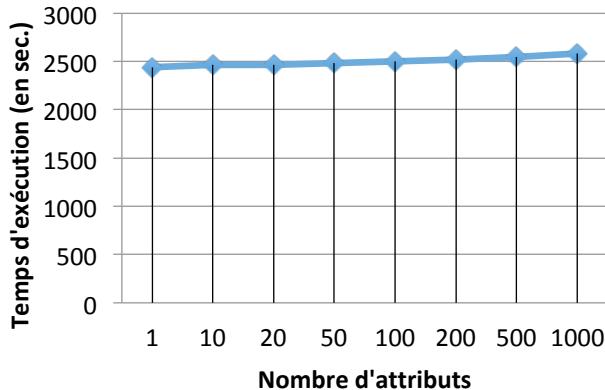
- < (Pittsburgh: C+, D-, WA+ | Providence: C+, D=, WD+ | Portland: C+, DD+ | Shreveport: D=, DD+, DA+ | ...) (Pittsburgh: C=, D-, DD- | Providence: D-, DD=, DA- | Portland: DD-, WD- | Shreveport: D=, DD- | ...) >
- Semaines du
- 08/08 22/08 05/09
- dégradation du trafic aérien (annulations et retards) après chaque ouragan avec un retour progressif à la normale la semaine suivante (impact sur la côte Est mais aussi sur certains aéroports de la côté Ouest)

Extraction de motifs dans un unique graphe dynamique attribué: Quelques résultats (2/2)

- Impact des données en entrée sur les performances et le nombre de solutions

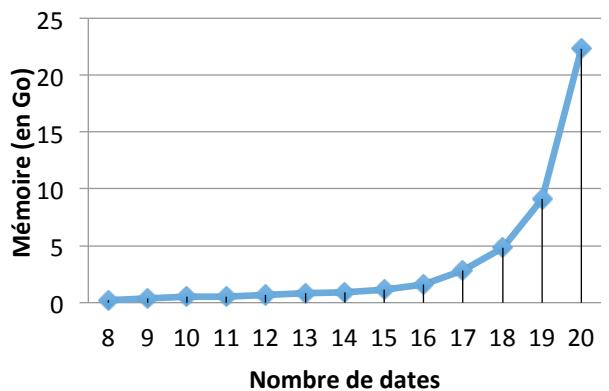
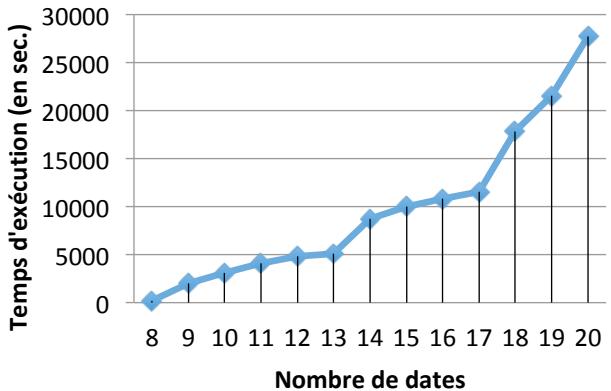
➤ Influence du nombre d'attributs

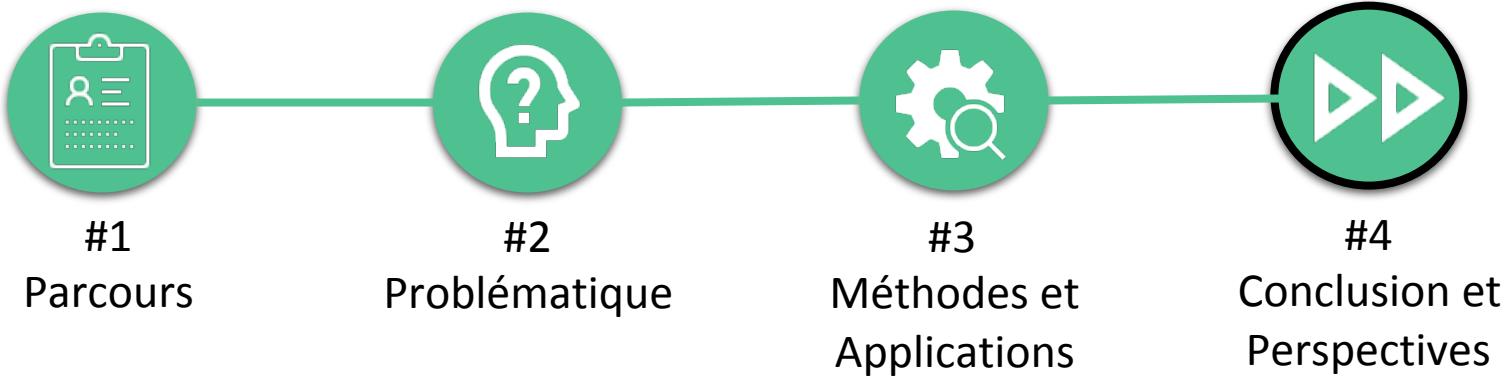
Peu d'impact sur les performances et le nombre de solutions



➤ Influence du nombre de dates/temps

Un impact fort sur les performances (notamment la mémoire)





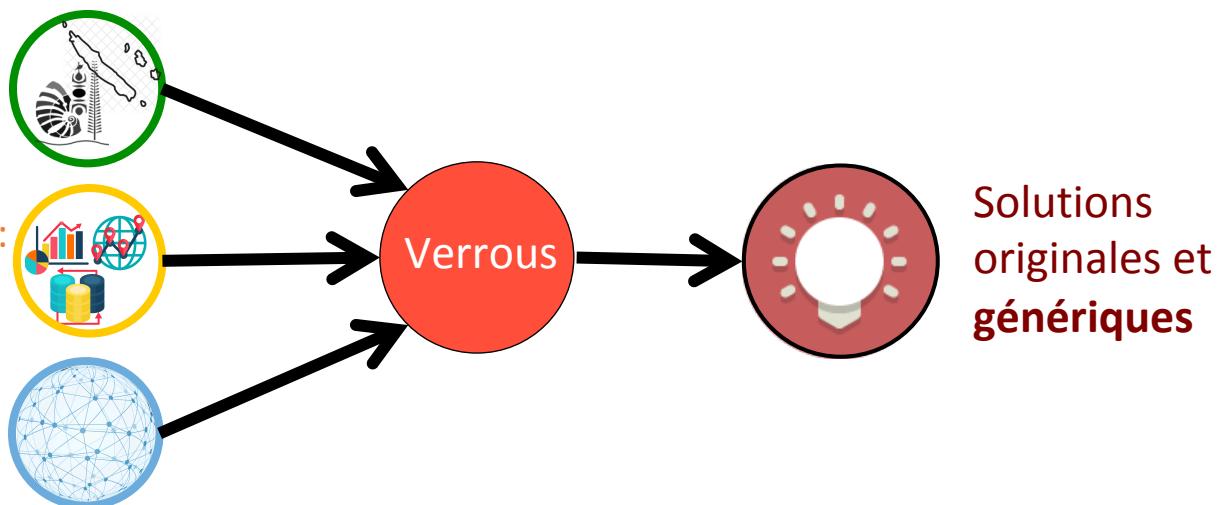
- **Problème étudié** : Extraction de motifs (régularités) dans des données spatio-temporelles complexes
- **Contributions de l'équipe** :
 - Des représentations des données et des domaines de **motifs** permettant de capturer de plus en plus d'**informations**
 - Des **méthodes et algorithmes** performants
 - Des **contraintes** statistiques, structurelles et **expertes** permettant d'améliorer la pertinence des résultats
 - Une **application** à des problématiques réelles en lien avec des experts
 - Un **transfert** vers des solutions logicielles adaptées aux utilisateurs

Contexte applicatif :

Suivi environnemental

Données spatio-temporelles variées :
SIG, images satellites, etc.

Travaux de la communauté :
Extraction de motifs séquentiels
et de graphes



- Amélioration de la pertinence des motifs extraits
 - Constat : beaucoup d'informations extraites peu pertinentes malgré la multitude de mesures proposées
 - Peu d'intégration de la connaissance des experts ou une intégration nécessitant un fort investissement de leur part
- Poursuivre les travaux sur l'exploitation des modèles existant dans le domaine d'application
 - Faire le lien avec les travaux visant à extraire des motifs "exceptionnels" par rapport à un modèle en entrée (*exceptional model mining*) [Duivesteijn et al., DMKD'16]
 - Extraction de motifs supervisée par des modèles statistiques
 - Des modèles plus complexes (notamment temporels) :
 - Modèles à base d'équations différentielles ordinaires (ODE)
 - Modèles à base d'agents

En cours : Projet SOSPADIS

Modélisation de la dynamique des habitats informels ("squats") à Fidji et au Vanuatu avec une intégration de modèles multi-agents et de techniques de science des données

- Meilleure intégration des données numériques dans les méthodes d'extraction de motifs

- Constat : nécessité de transformer les données numériques en intervalles avant de pouvoir les analyser ("discrétisation")
- Souvent une construction empirique avec une perte d'informations en sortie

➤ Construction dynamique d'intervalles pendant l'extraction, p.ex. [Srikant et Agrawal, SIGMOD'96] [Grosskreutz et Ruping, DMKD'09] [Kaytoue et al., IJCAI'11]

- Itemsets quantitatifs, sous-groupes, motifs intervalles

- Des problèmes de performance

• Etudier cette problématique dans le contexte des données spatio-temporelles

- Des données avec des spécificités (p.ex. notions de hiérarchie et de granularité)

➤ Exploiter ces spécificités pour améliorer le passage à l'échelle ?

- **Intégration et distribution de l'analyse dans les réseaux d'objets connectés (IoT)**
 - **Constat** : Explosion du nombre d'objets connectés, de la quantité de données générées par ces objets et des besoins en analyse
 - Difficile de rapatrier les données générées sur des serveurs pour les analyser (bande passante limitée)
 - Difficile d'intégrer toute l'analyse dans les objets (ressources limitées)
 - Etudier la distribution des traitements entre le *cloud* et les objets, et l'exploitation de nouvelles architectures matérielles (p.ex. puce Edge TPU, NVIDIA Jetson Nano)
- Plus généralement, ouverture vers d'autres problématiques et d'autres domaines d'application
 - **La classification supervisée**
 - Classification à partir de règles d'association dans des séries temporelles multi-variées
 - La santé et l'éducation

En cours :

Partenariat avec une startup développant des instruments de suivi connectés

Partenariat en enseignement avec une association développant des mains robotisées pour des enfants en situation de handicap

Merci pour votre attention

