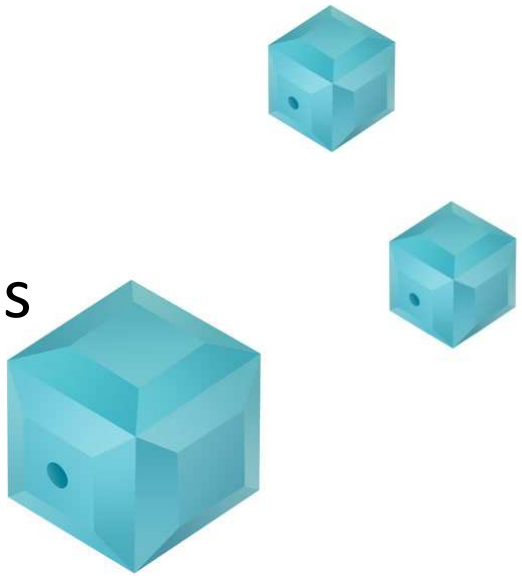
Vers l'OLAP sémantique pour l'analyse en ligne des données complexes

Sabine Loudcher

Habilitation à Diriger des Recherches

Laboratoire ERIC, IUT Lumière

Université Lyon 2



29 juin 2011

ERIC LABO EQUIPE DE RECHERCHE EN INGENIERIE DES CONNAISSANCES

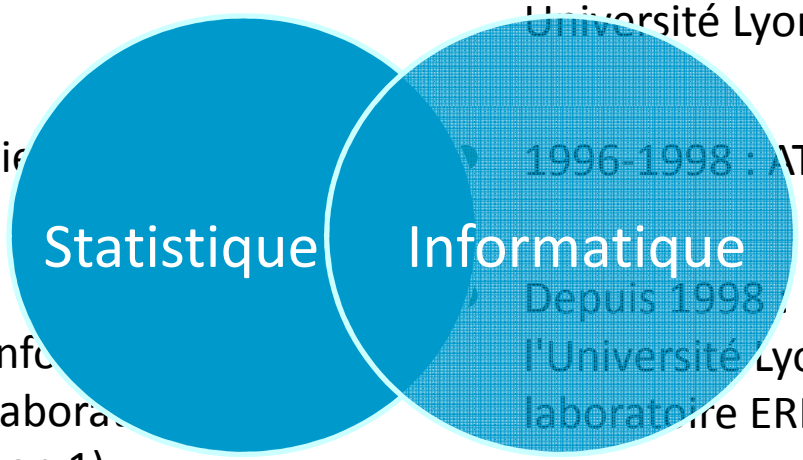
UNIVERSITÉ LUMIÈRE LYON 2 UNIVERSITÉ DE LYON

UNIVERSITÉ DE LYON

Parcours

Diplômes

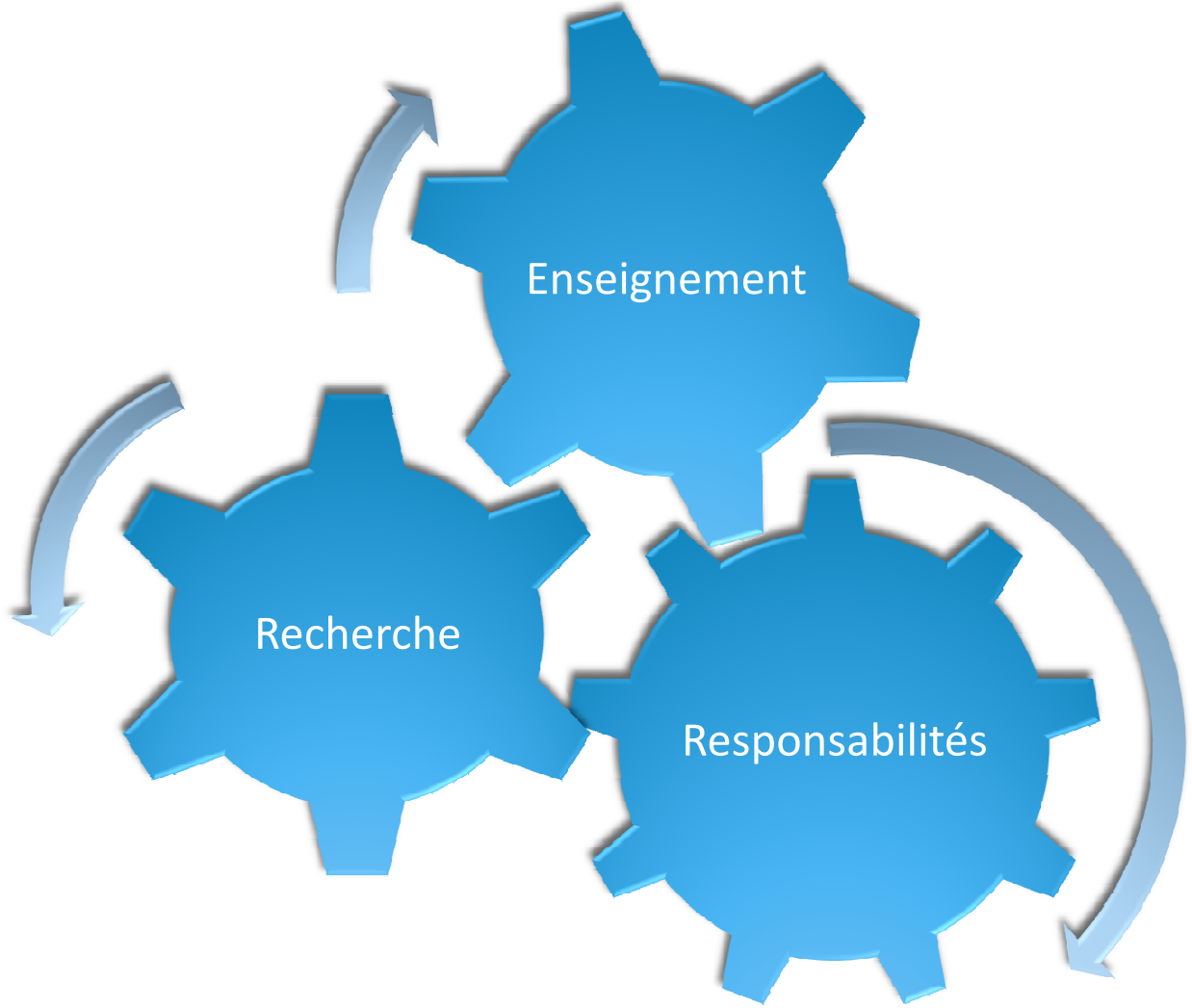
- 1992 : DESS de Statistique et Informatique Socio-Economiques, Université Lyon 2
- 1994 : DEA d'Ingéni
  Université Lyon 1 –
- 1996 : Doctorat d'Inf
  Université Lyon 1, labora
  (URA 934 CNRS - Lyon 1),
  Pr D.A. ZIGHED (directeur de thèse)



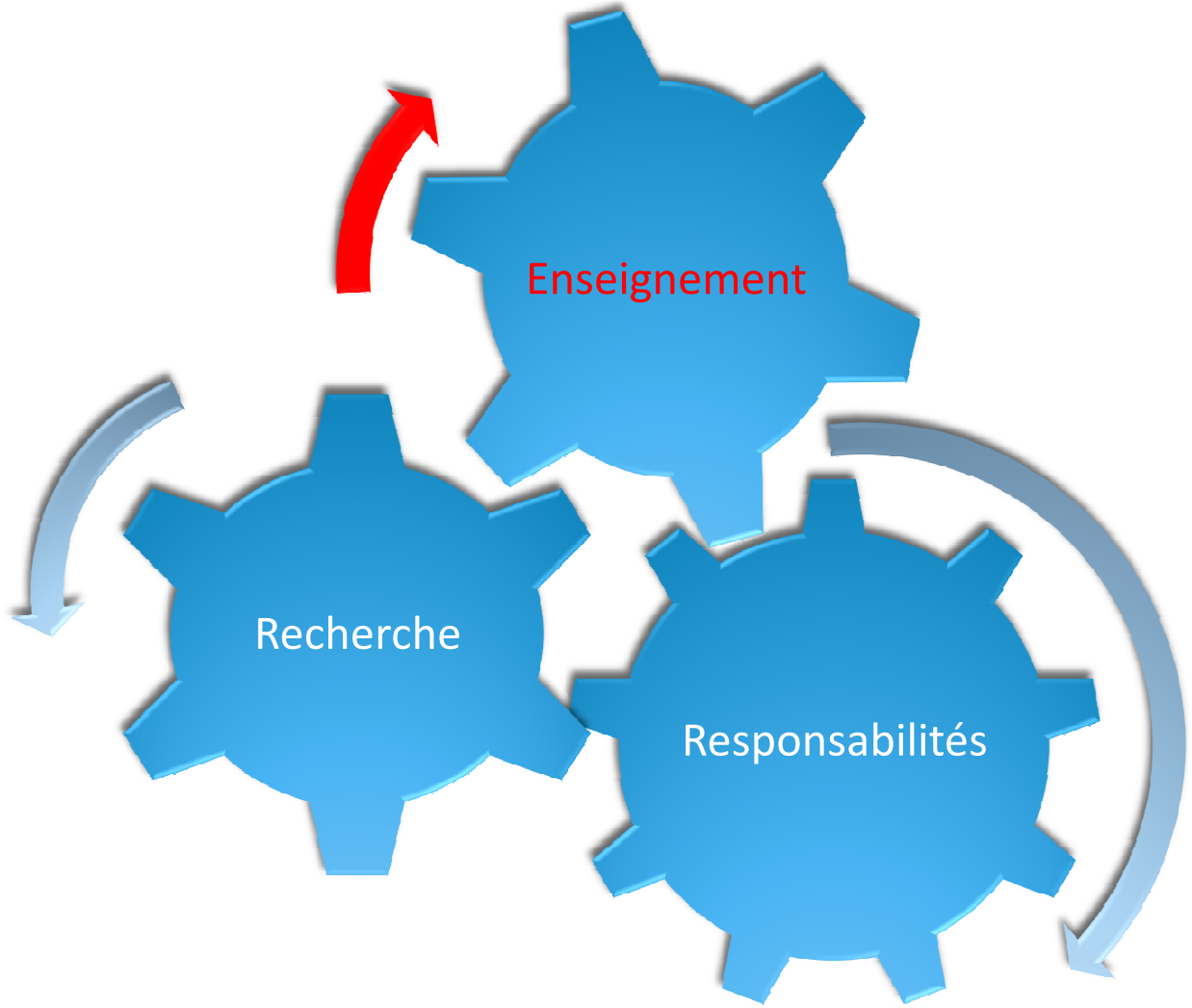
Carrière universitaire

- 1994-1996 : Doctorant avec une bourse de docteur-ingénieur du CNRS,
  Université Lyon 1
- 1996-1998 : ATER, Université Lyon 3
- Depuis 1998 : Maître de Conférences à l'Université Lyon 2 (IUT Lumière, laboratoire ERIC)

Parcours



Parcours



Enseignement

Maître de conférences

2011

Vacataire, doctorant, ATER

1998

1992

IUT LUMIERE LYON 2 – Le choix de l'Alternance

Département d'Informatique et de Statistique – UNIVERSITÉ LUMIÈRE LYON 2 – DIS – Faculté de Sciences Économiques et de Gestion

dkm²

Erasmus Mundus master course in **Data Mining and Knowledge Management** a european master

STATISTIQUE ET TRAITEMENT INFORMATIQUE DES DONNEES – STID

Chargé d'Etudes Statistiques – CESTAT

Probabilités
Algorithmique et programm
Bases de données et tableur
Analyse de données
Apprentissage automatique
800 heures

essi ensional Data Analysis
année
Systèmes d'information décisionnels
Master recherche et professionnel 2ème année
Logiciels statistiques
Méthodologie de rédaction de mémoire
250 à 300 heures par an

Encadrement pédagogique

Maître de conférences

2011

1998

1992

IUT LUMIERE LYON 2
Le choix de l'Alternance

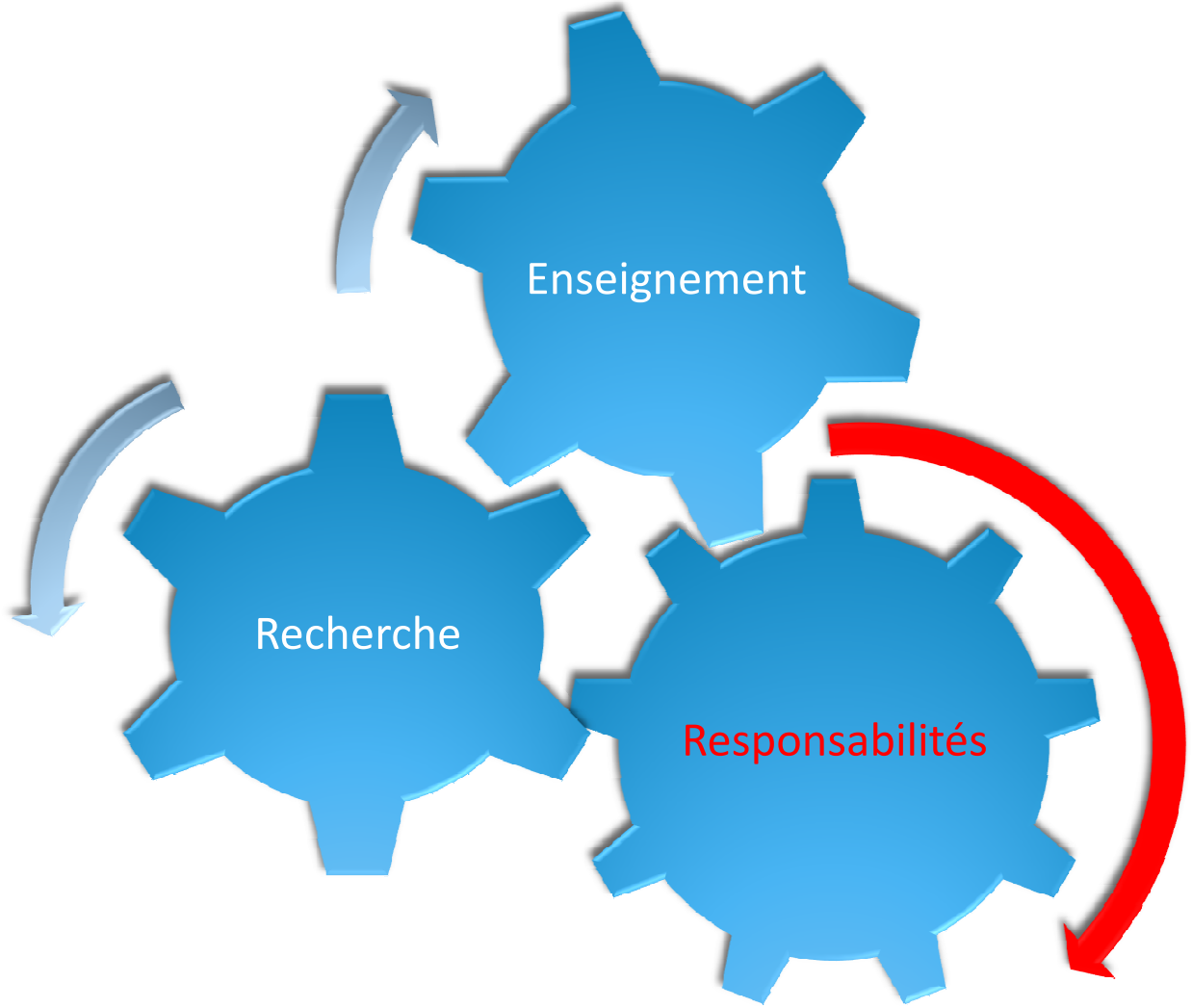
Département d'Informatique et de Statistique
UNIVERSITÉ LUMIÈRE Lyon 2
DIS

Encadrement d'étudiants en stage en entreprise

Etudiants en alternance
Accompagnement en entreprise
Tuteur enseignant

Encadrement de projets pédagogiques

Parcours



Implication universitaire

Responsabilités pédagogiques, administratives et institutionnelles

Chef du département STID

Chargée de mission pour la direction de l'IUT

Commission de spécialistes, groupe d'experts, jurys d'IGE

Directeur adjoint du laboratoire ERIC

1998 **2003** **2011**

Chef du département STID (1998-2002)

Création puis direction du département STID

- Définition des **orientations stratégiques** et pédagogiques du diplôme
- Mise en place d'une **pédagogie de l'alternance** efficace et adaptée
- Relations avec les **milieux professionnels et** les **partenaires institutionnels**
- Réflexion sur les **débouchés** de la formation
- Première **évaluation quadriennale**
- **Gestion financière** et **administrative** du département

Chargée de mission pour l'IUT (2003-2010)

- **Responsable du projet «Observatoire Etudiants »**
  - Mise en place du projet
  - Conception en ligne de tableaux de bord sur le recrutement, la formation et le devenir des étudiants
  - Encadrement de l'équipe de développement

- **Pilotage des enquêtes sur l'insertion professionnelle des apprentis dans l'enseignement supérieur de la région Rhône-Alpes**
  - Collaboration étroite avec le comité régional Forma-Sup de l'apprentissage et les rectorats
  - Expertise statistique
  - Encadrement de l'équipe de développement

- **Représentation de l'université Lyon 2 dans les instances régionales pilotant l'apprentissage dans l'enseignement supérieur**
  - Conseil d'Administration du CFA Forma-Sup ARL
  - Comité régional Forma-Sup

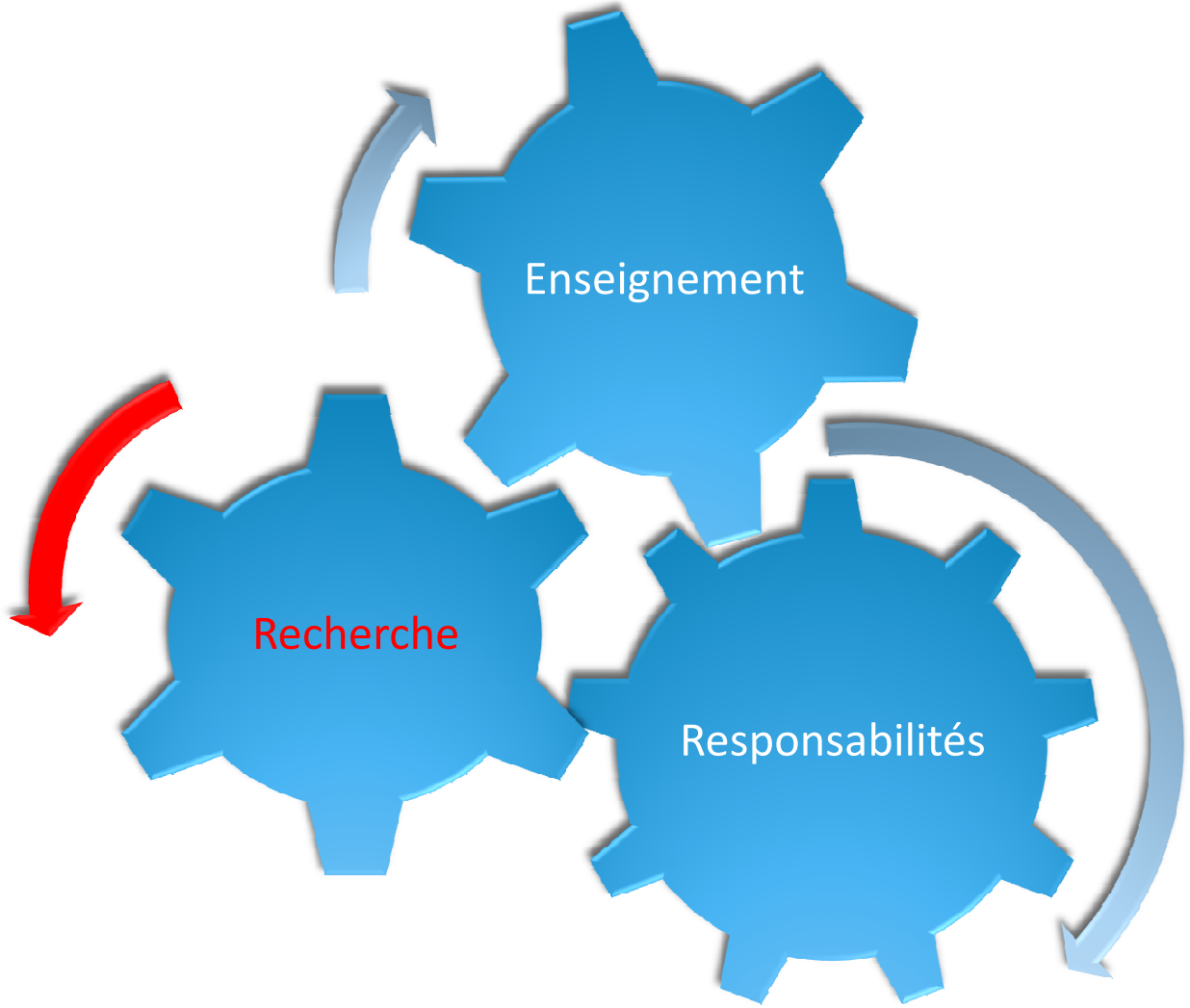
Directeur adjoint du laboratoire (2003-)

- **Depuis 2003, sous la direction de Nicolas Nicoloyannis puis de Djamel Zighed**

- Gestion financière (élaboration et suivi du budget)
- Gestion des ressources humaines
- Communication (site Web, plaquettes, …)
- Relations avec les services internes de l'université
- Fonctionnement administratif et quotidien du laboratoire
- Préparation des évaluations du laboratoire
- Préparation des décisions discutées en conseil de direction et votées en conseil de laboratoire

- **Fonctions et missions accrues et renforcées depuis 2010**

Parcours



Thématique scientifique

- Informatique décisionnelle, entrepôts de données et analyse en ligne
  - Collecter, organiser, stocker et analyser l'information
  - Aider la prise de décision
- Avènement des données complexes
  - Données multi-format, multi-structure, multi-source, multi-modal, multi-version, riches en sémantique
- Remise en cause du processus d'entreposage et d'analyse
- Nouveaux problèmes de recherche : intégration, stockage, modélisation et analyse des données complexes

Positionnement des travaux

Contexte

- OLAP et données complexes (DC)

- Vocation de l'analyse en ligne (OLAP)
  - Analyse interactive et multidimensionnelle des données de l'entrepôt
  - Agrégation des données pour résumer, explorer, visualiser
  - Représentation sous forme de cube et manipulation avec des opérateurs

Problèmes

- Pas d'outils automatiques
- Pas d'extraction de connaissances
- Opérateurs OLAP inadaptés pour les DC
- Comment agréger les DC ?
- Comment visualiser les DC ?
- Comment prendre en compte la sémantique contenue dans les DC ?

Problématique de recherche

Analyse en ligne des données complexes

Enrichir les possibilités de l'OLAP

- Explication
- Prédiction

Créer une analyse adaptée aux DC

- Visualisation
- Régions intéressantes

Prendre en compte la sémantique

- Agrégation sémantique

Contributions

Originalité

- Combinaison de l'OLAP et de la fouille de données
  - Analyse plus intéressante
  - Question non triviale

- Combinaison de l'OLAP et de la recherche d'information

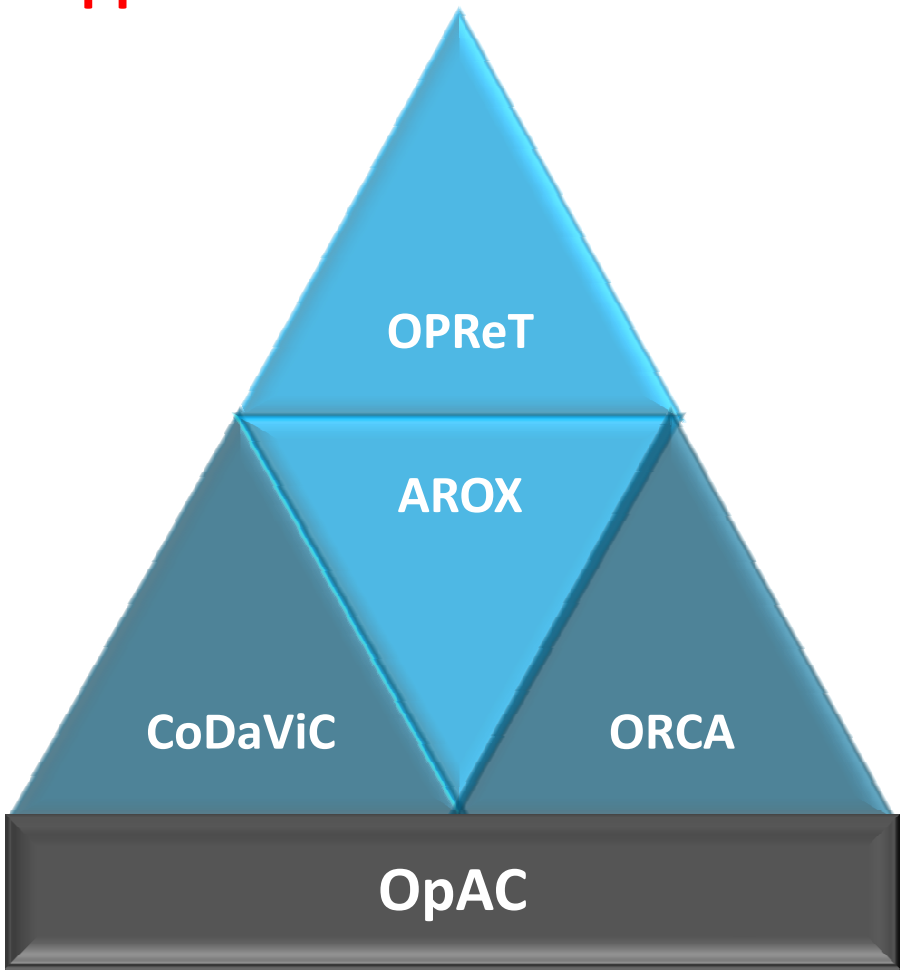
- Analyse **explicative** par recherche guidée de règles d'association
- Analyse **prédictive** avec les arbres de régression

- **Visualisation** dans un cube de DC avec une méthode factorielle
- Détection de **régions intéressantes** avec une méthode factorielle

- **Agrégation sémantique** avec une méthode de classification

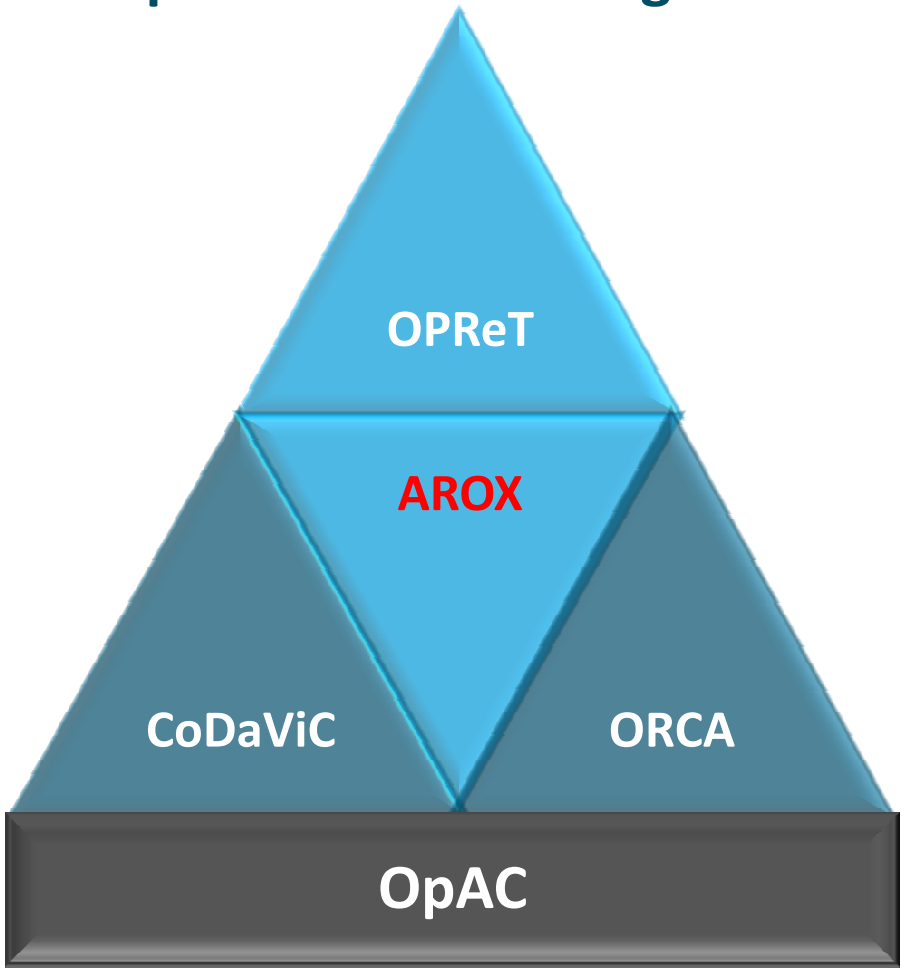
Contributions

Cinq nouvelles approches



Contribution

Analyse explicative par une recherche guidée de règles d'association



Analyse explicative par une recherche guidée de règles d'association

Problème

CA (en euro)	T1	T2	T3	T4
Imprimante	9400	10000	12600	10500
MP3	20500	13700	54400	21000
PC	13100	14600	15200	12300
PC portable	11400	12000	28000	10000

Pourquoi les ventes de lecteurs MP3 sont-elles particulièrement élevées au 3ème trimestre ?

CA		Juin	Juillet	Août
(en euro)	Jeunes	9300	24300	19100
MP3	Adultes	1200	600	1600
	Agés		300	

Les mois des mois de juillet, août et les jeunes consommateurs sont associés aux ventes élevées de lecteurs MP3

- Pas d'outils OLAP automatiques pour expliquer les relations et les associations
- Besoin d'une nouvelle possibilité d'analyse : l'explication
- **Comment expliquer automatiquement des phénomènes ? Comment détecter des associations ?**

Analyse explicative par une recherche guidée de règles d'association

Motivation

- **Utiliser le principe des règles d'association**
  - Technique de fouille de données avec le même objectif
  - Structure multidimensionnelle, un contexte favorable
- **Contribution** : AROX (*Association Rules Operator for eXplication*)
- **Positionnement**
  - Travaux de (Kamber 1997), (Zhu 1998), (Imielinski 2002), (Tjioe et Taniar 2005)
  - Fouille guidée par une méta-règle
  - Règles inter-dimensionnelles
  - Recherche des motifs fréquents et des règles dans la structure multidimensionnelle
  - Modification de la définition du support et de la confiance pour l'adapter à l'OLAP

Analyse explicative par une recherche guidée de règles d'association

Principe

- Support et confiance basés sur la mesure

$$R: Continent = Amérique \wedge Année = 2009 \rightarrow Produit = MP3$$

Nb de ventes	2009 Amérique	2009 Europe	2010 Amérique	2010 Europe
PC	1200	800	950	500
PC portable	2500	2400	2800	3010
MP3	11600	5900	11400	9100

Chiffre d'affaires	2009 Amérique	2009 Europe	2010 Amérique	2010 Europe
PC	60000€	33000€	28000€	10000€
PC portable	500000€	560700€	420000€	544000€
MP3	116000€	118000€	57000€	41000€

Définition classique : comptage des faits

$$Supp(R) = \frac{NB(Amérique, MP3, 2009)}{NB(All, All, All)}$$

$$Conf(R) = \frac{NB(Amérique, MP3, 2009)}{NB(Amérique, All, 2009)}$$

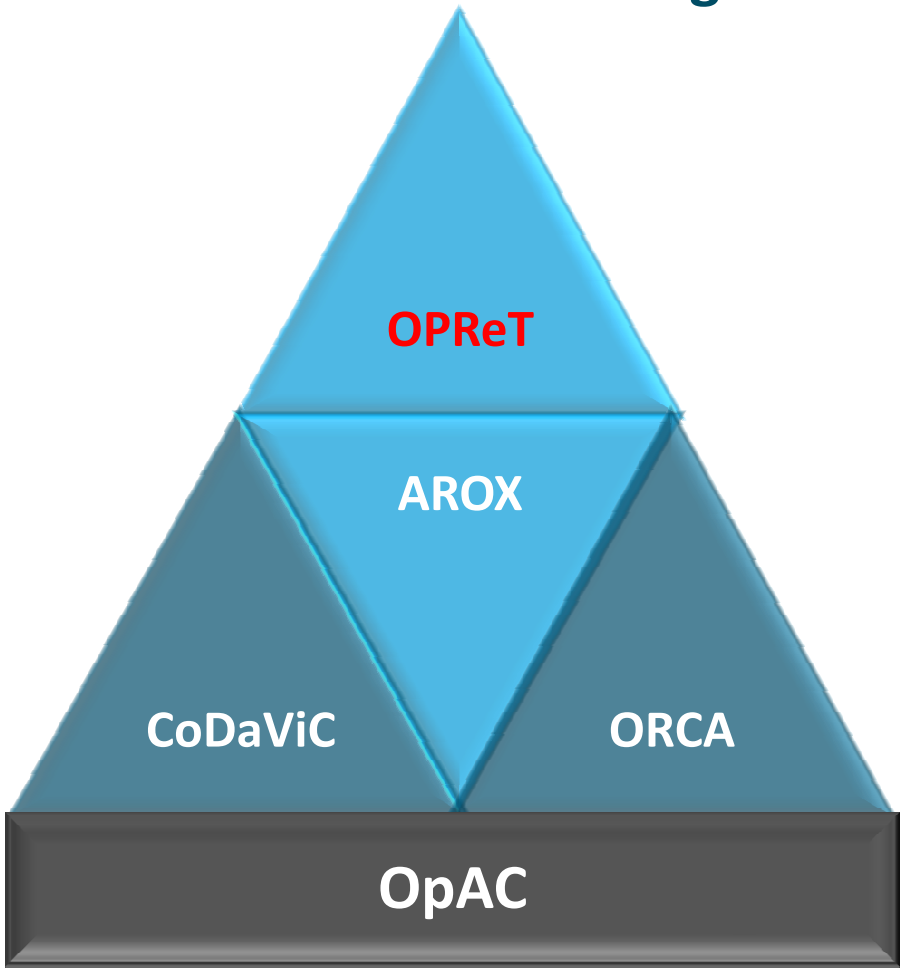
Nouvelle définition : avec la mesure

$$Supp(R) = \frac{SUM_{CA}(Amérique, MP3, 2009)}{SUM_{CA}(All, All, All)}$$

$$Conf(R) = \frac{SUM_{CA}(Amérique, MP3, 2009)}{SUM_{CA}(Amérique, All, 2009)}$$

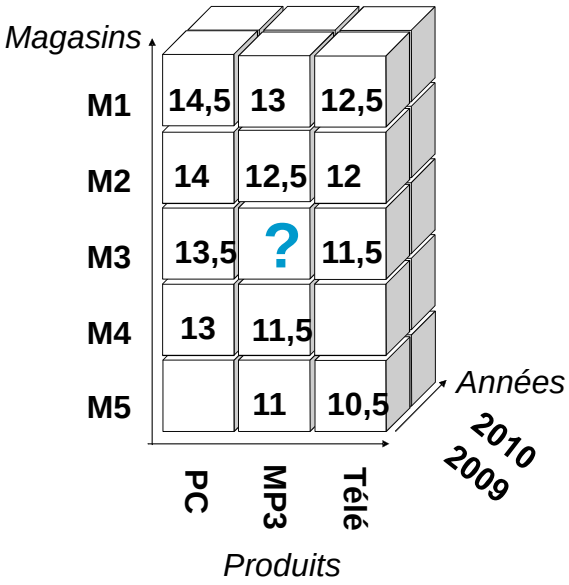
Contribution

Analyse **prédictive** avec les arbres de régression



Analyse prédictive avec les arbres de régression

Problème



- Besoin d'analyse de l'utilisateur : « qu'est ce qui se passe si … ? »
- Comment, à partir des cellules pleines voisines, donner une valeur à une cellule vide désignée par l'utilisateur ?
- Pas d'opérateurs OLAP classiques, nouveau besoin d'analyse en ligne : la prédiction
- **Comment intégrer la prédiction dans l'OLAP ?**

Analyse prédictive avec les arbres de régression

Motivation

- Dans le cadre du ***What If Analysis*** (Golfarelli 2006)
- **Couplage** entre l'OLAP et la fouille de données pour prédire la mesure
- **Positionnement**
  - Travaux de (Han et S. Cheng 1998), (Sarawagi 1998), (BC. Chen 2005, 2006), (Y. Chen et J. Pei 2001, 2006), (Palpanas 2001, 2005)
  - Prédire la valeur d'une mesure pour un nouveau fait et compléter le cube
  - Placer l'utilisateur au centre ; donner des indicateurs de qualité
  - Fournir un modèle **utilisable** dans l'OLAP, facilement **interprétable**, **sans hypothèse**
  - Intégrer une démarche complète d'apprentissage supervisé

- **Contribution** : OPReT (*Online Prediction by Regression Tree*)

Analyse prédictive avec les arbres de régression

Principe

1. Contexte d'analyse

Sous-cube de données

2. Modèle de prédiction

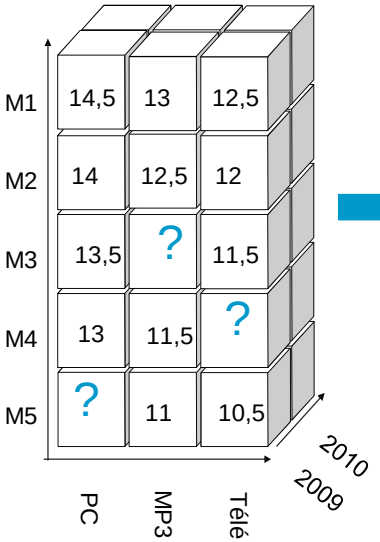
Construction
Validation
Taux d'erreur moyen
Réduction de l'erreur

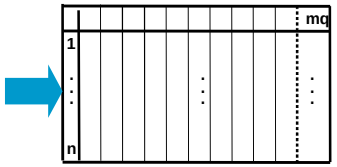
3. Interprétation du modèle

Règles de décision
Indicateurs (support, écart-type)

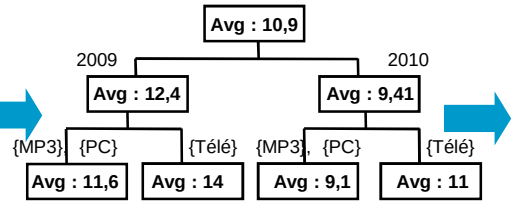
4. Prédiction OLAP

Choix des cellules
Valeurs prédites intégrées
Intégration visuelle

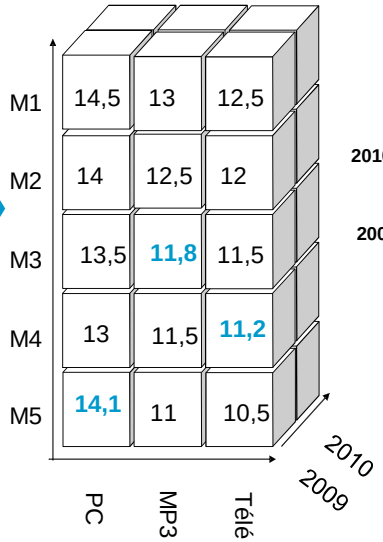


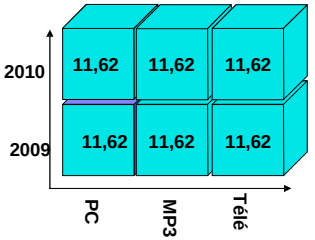


70% des faits pour l'apprentissage
30% pour le test



R (X → Y; S; σ)





Calcul des nouveaux agrégats lors d'un forage vers le haut

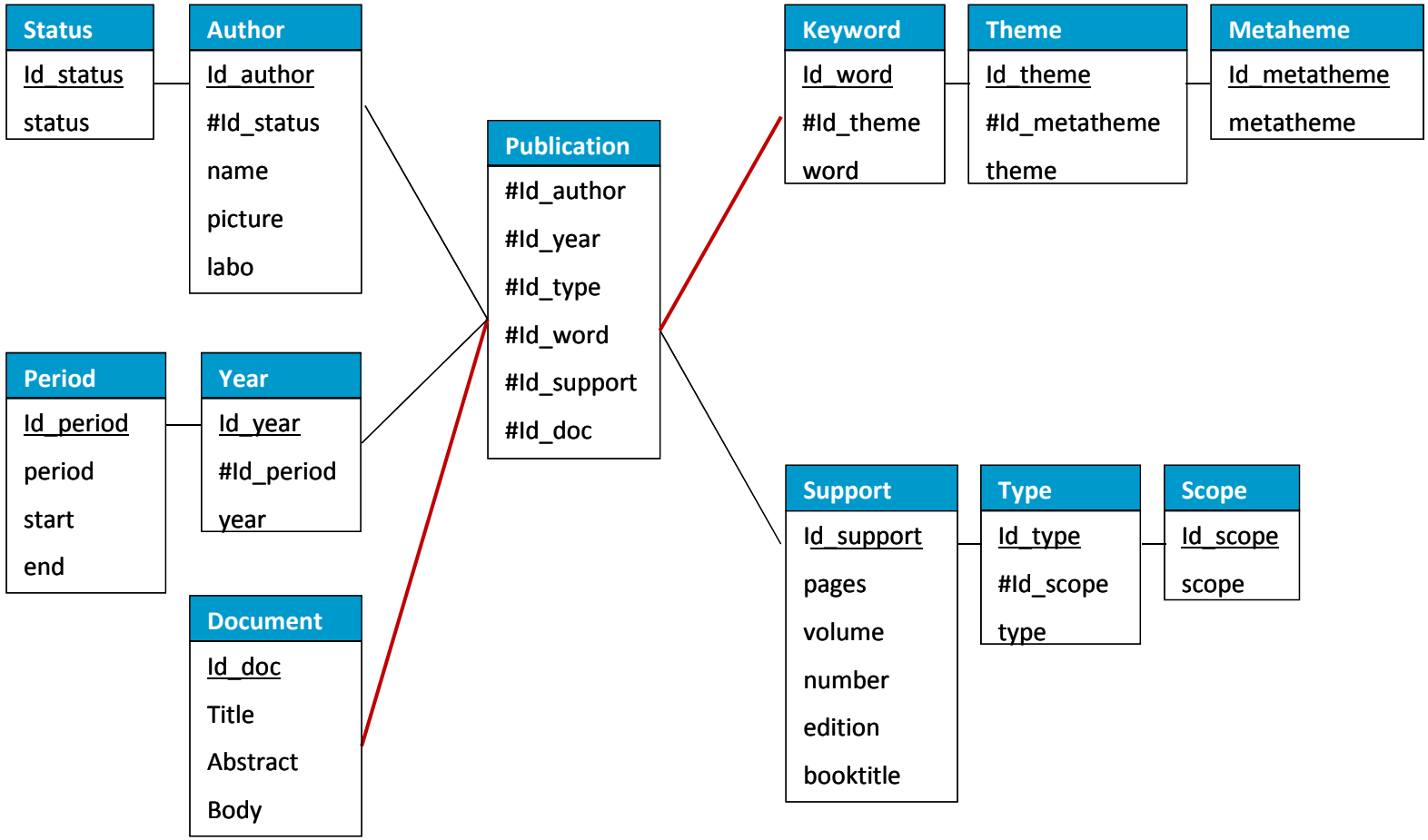
Analyse en ligne des données complexes

- **Avènement des données complexes**

- **Verrous scientifiques** posés par les données complexes dans l'analyse en ligne
  - Visualiser l'information contenue dans les cubes de DC
  - Organiser les cubes de DC pour améliorer la visualisation et détecter des régions intéressantes
  - Agréger des données complexes
  - Prendre en compte le contenu sémantique des données
- **Exemple de l'analyse des publications scientifiques**
  - Publications = données complexes, entités sémantiques
  - Publication = {auteurs, titre, document, date, support, …}

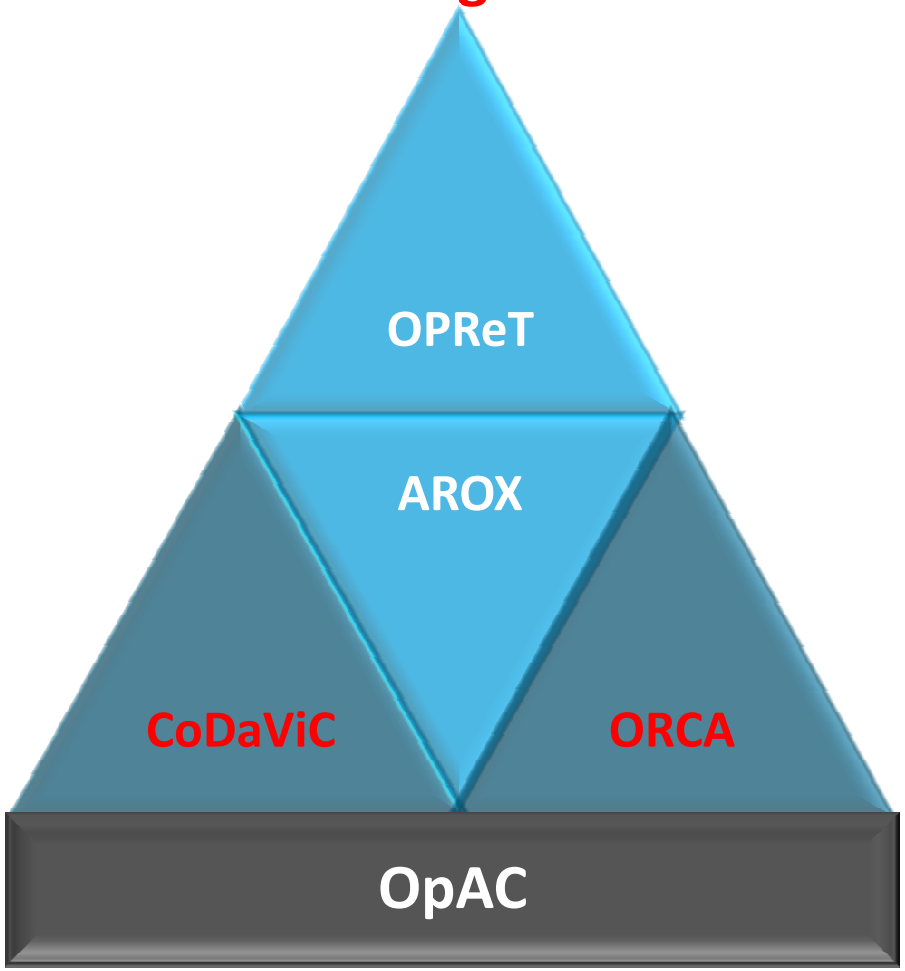
Analyse en ligne des données complexes

Modélisation multidimensionnelle des publications



Contribution

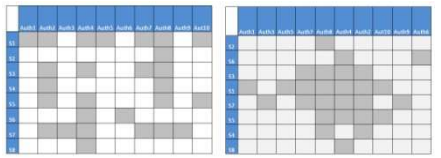
Visualisation et détection de régions intéressantes



Visualisation et régions intéressantes

Problèmes

- **Pas d'outils de visualisation OLAP adaptés aux données complexes**
  - Les faits = des données comportant du texte, des images, …
  - Pas toujours une mesure ou pas de mesure numérique
- **Exploration OLAP manuelle et intuitive du cube**
  - Navigation parfois longue et non triviale
  - Eparsité des cubes de données complexes
  - Modalités des dimensions ordonnées selon un ordre pré-établi



- **Comment représenter l'information contenue dans un cube de DC ?**
- **Comment organiser le cube de DC pour détecter des régions intéressantes ?**

Détection de régions intéressantes

Problème

	Auth1	Auth2	Auth3	Auth4	Auth5	Auth6	Auth7	Auth8	Auth9	Aut10
S1										
S2										
S3										
S4										
S5										
S6										
S7										
S8										

	Auth1	Auth3	Auth5	Auth7	Auth8	Auth4	Auth2	Aut10	Auth9	Auth6
S2										
S6										
S3										
S1										
S7										
S5										
S4										
S8										

Visualisation et régions intéressantes

Motivations

- **Positionnement**
  - Peu de travaux : (Ordonez 2009), (Aouiche & Lemire 2008), (Morin 2007), (Sureau &Venturini 2009)
  - Pas de réduction de l'éparsité, pas de compression
- **Visualiser en ligne** l'information contenue dans un cube
  - Visualisation adaptée aux DC, prise en compte du contenu sémantique
  - Aucune hypothèse sur les données
- **Guider** l'utilisateur vers les régions intéressantes
  - Réorganiser intelligemment les dimensions et modalités du cube
  - Détecter les faits intéressants ainsi que leurs liens
- **Deux contributions**
  - CoDaViC (*Complex Data Visualization by Correspondences*)
  - ORCA (*Operator for Reorganization by multiple Correspondence Analysis*)

Visualisation et régions intéressantes

Principe

- **Deux méthodes factorielles**
  - Analyse des correspondances (AFC)
  - Analyse des correspondances multiples (ACM)
- **Cube de données complexes**
  - Au minimum dénombrement des faits
  - Tableaux de contingence
- **Une méthode factorielle pour**
  - Réduire l'espace de représentation
  - Produire des axes factoriels (nouvelles dimensions)
  - Créer un nouvel espace de représentation des faits
  - Visualiser l'information dans le cube OLAP
  - Mettre en évidence des points de vue intéressants pour l'analyse

Visualisation avec une méthode factorielle

Principe

1. Contexte d'analyse

Sous-cube de données

2. Tableau de contingence

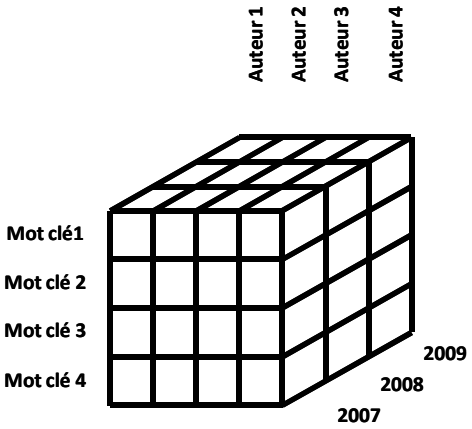
Opérateurs OLAP

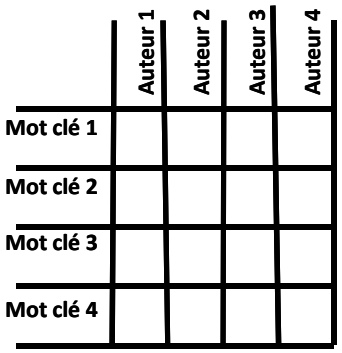
3. Analyse factorielle

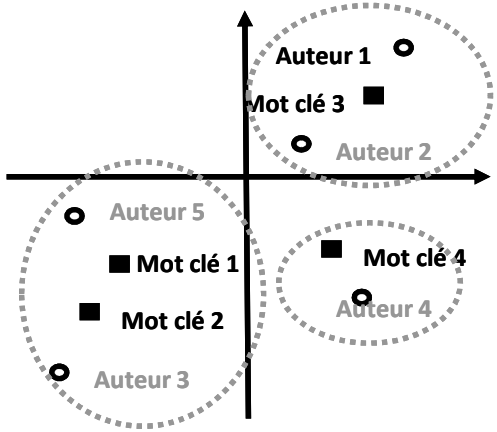
Axes factoriels
Projection des faits
Interprétation des proximités

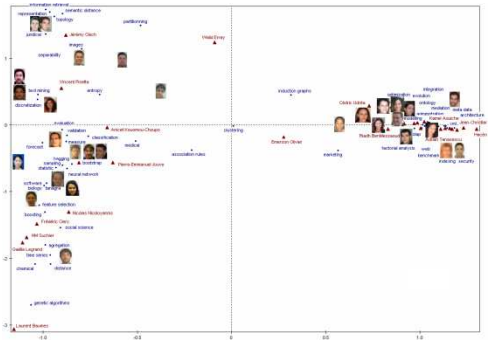
4. Visualisation

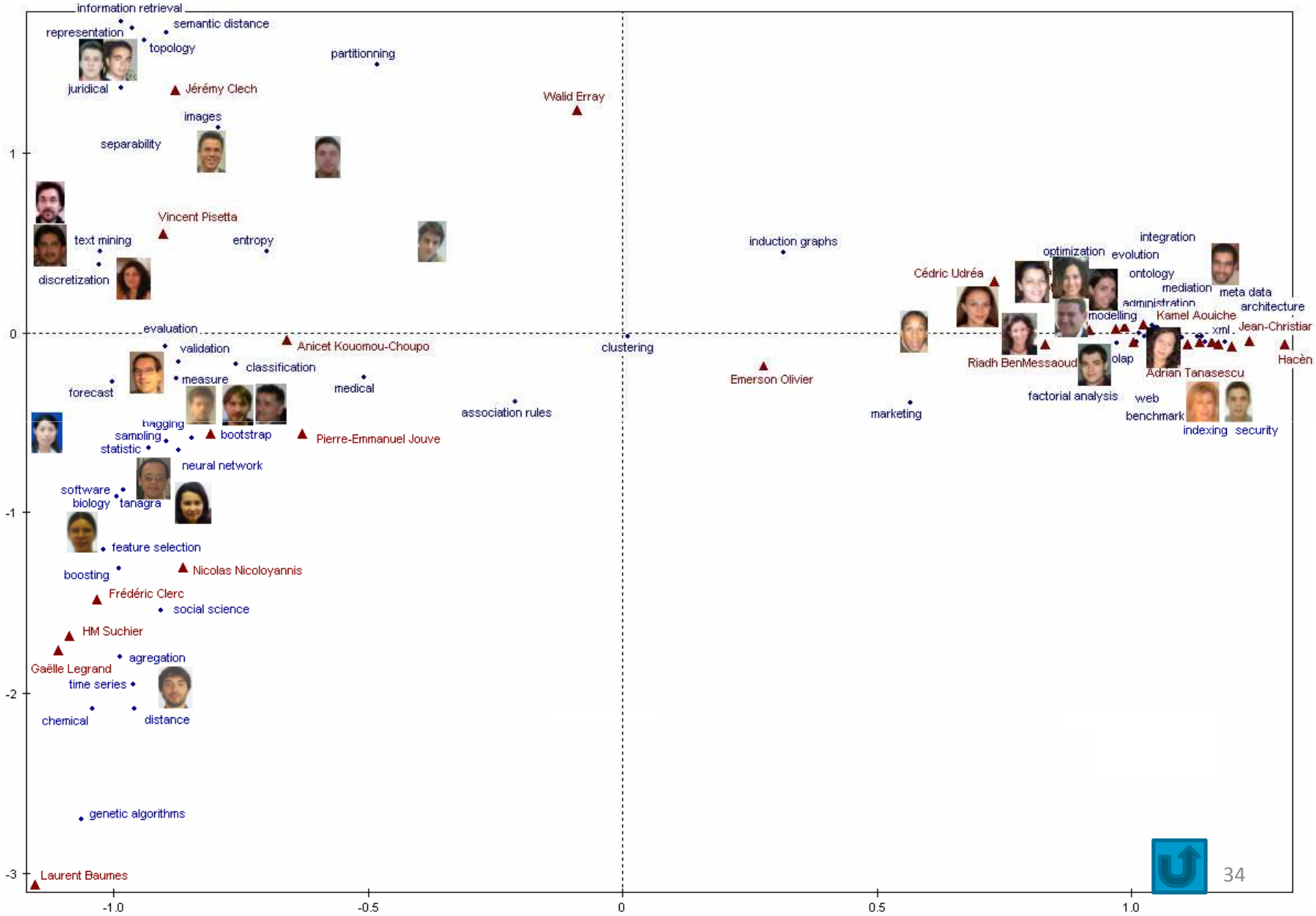
Synthèse graphique
Drill down possible





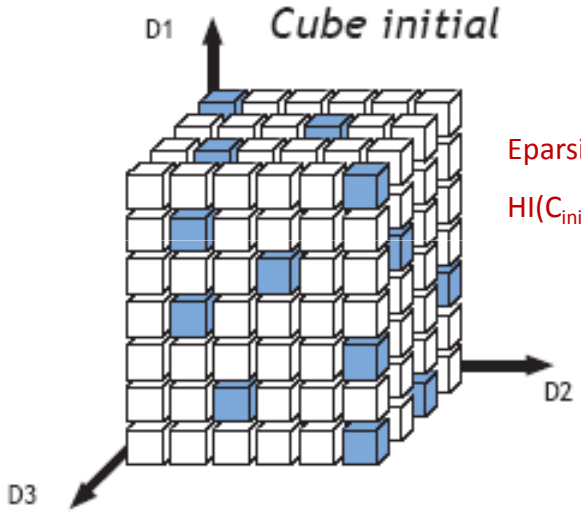





information retrieval
semantic distance
representation
topology
partitionning
juridical
Jérémy Clech
Walid Erray
images
separability
Vincent Pisetta
text mining
entropy
induction graphs
integration
optimization
evolution
ontology
Cédric Udréa
mediation
meta data
discretization
modelling
Kamel Aouiche
architecture
xml
Jean-Christiar
evaluation
validation
Anicet Kouomou-Choupo
clustering
classification
measure
Emerson Olivier
olap
Riadh BenMessaoud
Adrian Tanasescu
Hacèn
forecast
medical
factorial analysis
association rules
marketing
web
benchmark
bagging
sampling
bootstrap
Pierre-Emmanuel Jouve
indexing
security
statistic
neural network
software
biology
tanagra
feature selection
boosting
Nicolas Nicoloyannis
Frédéric Clerc
social science
HM Suchier
agregation
Gaëlle Legrand
time series
chemical
distance
genetic algorithms
Laurent Baumes
1
0
-1
-2
-3
-1.0
-0.5
0
0.5
1.0

Détection de régions intéressantes

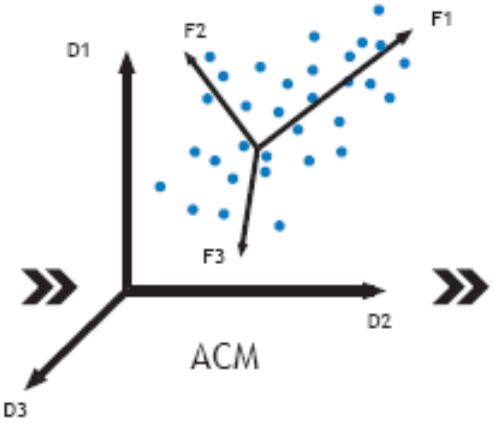
Principe

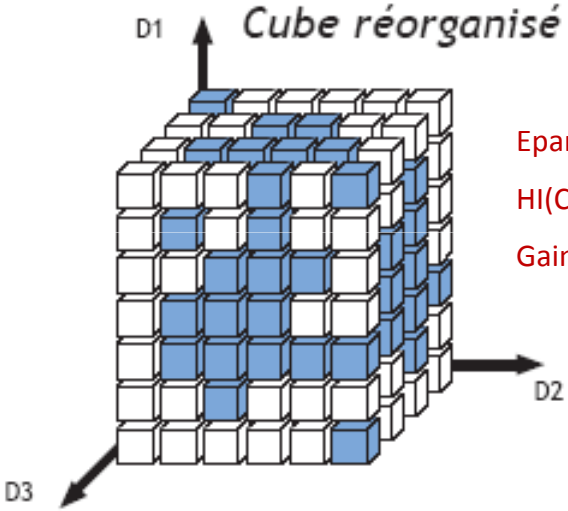


Eparsité = 63%

$HI(C_{ini})$ = 14%

	D1	D2	…	D3	Measure
1	-	-		-	-
2	-	-		-	-
.	-	-		-	-
.	-	-		-	-
.	-	-		-	-
N	-	-		-	-





Eparsité = 63%

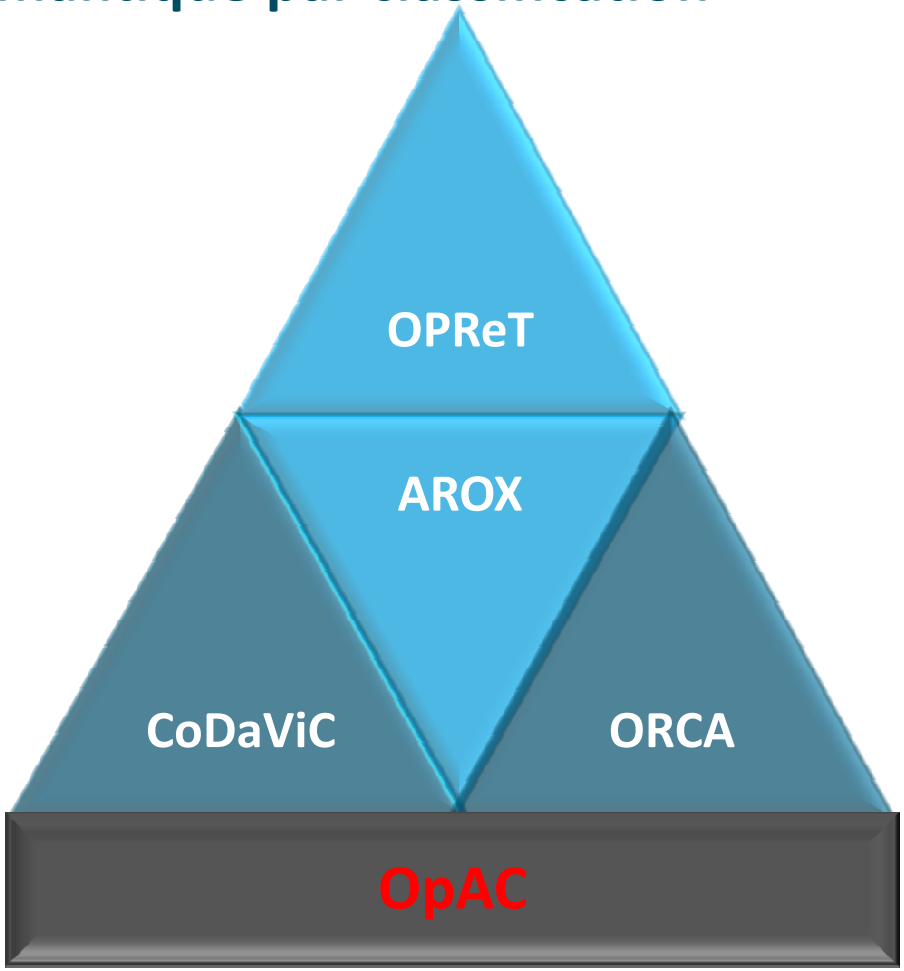
$HI(C_{arr})$ = 25%

Gain = 78%

	F1	F2	…	F3	Measure
1	-	-		-	-
2	-	-		-	-
.	-	-		-	-
.	-	-		-	-
.	-	-		-	-
N	-	-		-	-

Contribution

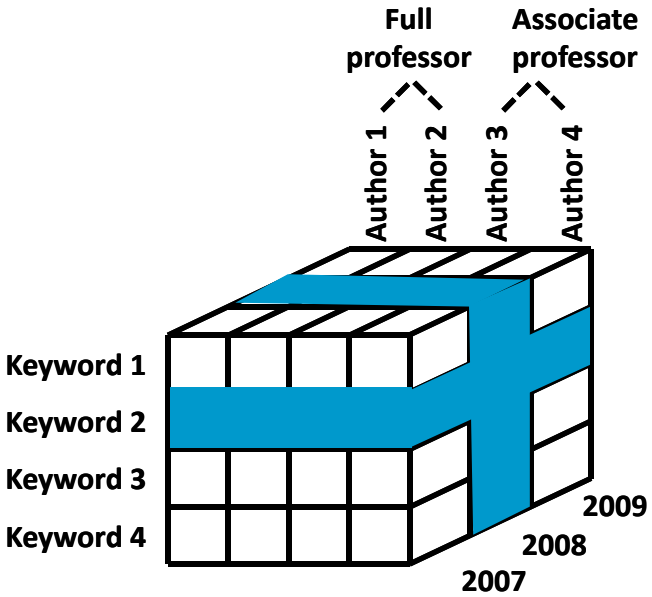
Agrégation sémantique par classification

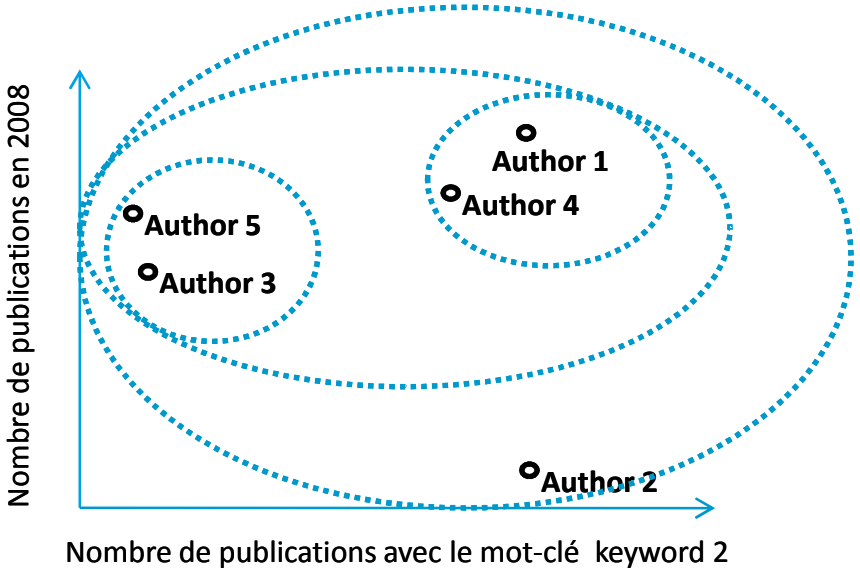


Agrégation sémantique par classification

Problème

- Classiquement, hiérarchies de dimensions fixées par l'expert
- Pas d'agrégation sémantique
- Pas d'agrégation adaptée aux DC





Agrégation sémantique par classification

Motivation

- **Agrégation sémantique**
  - Agrégation des faits selon leur proximité
  - Exploitation des mesures pour l'agrégation
- **Création** d'une hiérarchie de dimension
- **Classification Ascendante Hiérarchique** (CAH)
  - Hiérarchie de partitions = hiérarchie d'une dimension
  - Opérations *roll-up* et *drill-down* possibles
  - Stratégie ascendante vs descendante

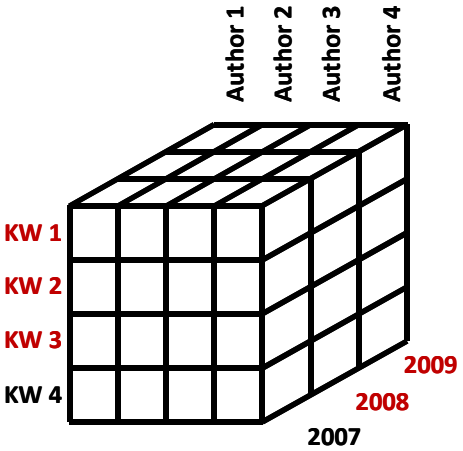
- **Contribution** : OpAC (*Operator for Aggregation by Clustering*)

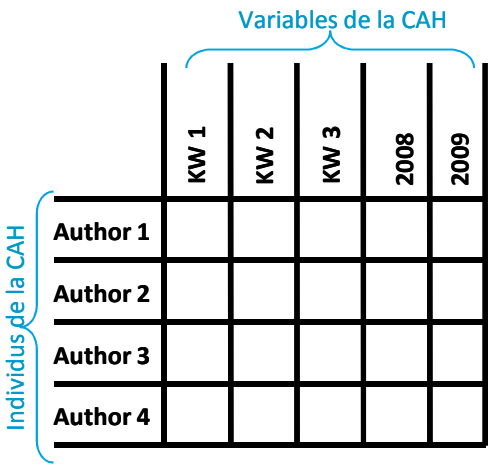
Agrégation sémantique par classification

Principe

1. Individus et variables de la classification

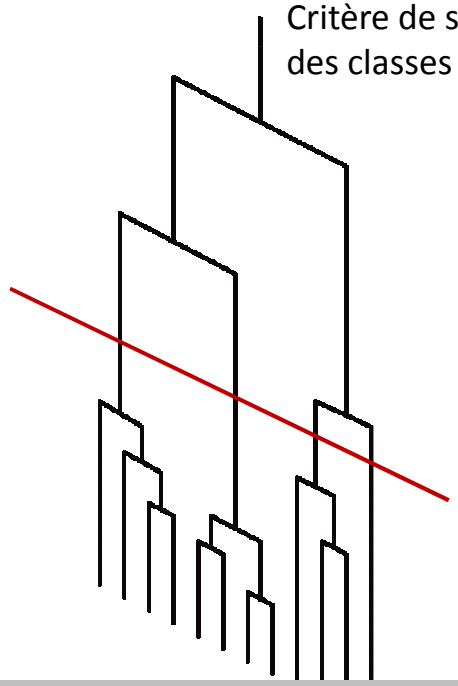
Choix des individus et des variables
Règles à respecter





2. Classification

Ascendante hiérarchique



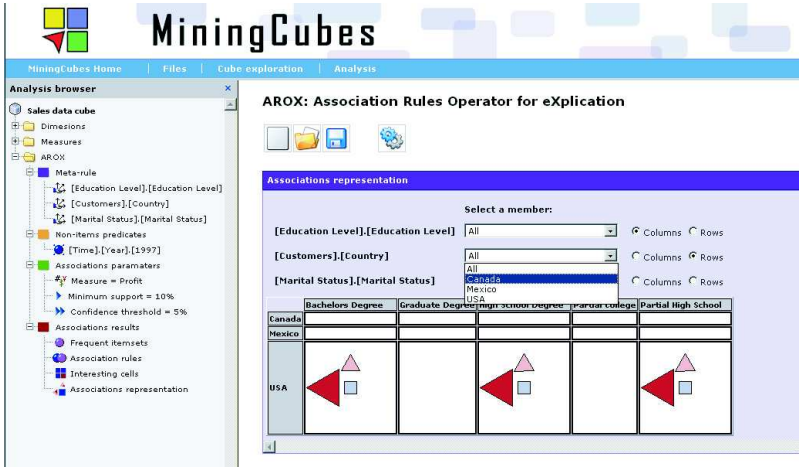
3. Evaluation des agrégats

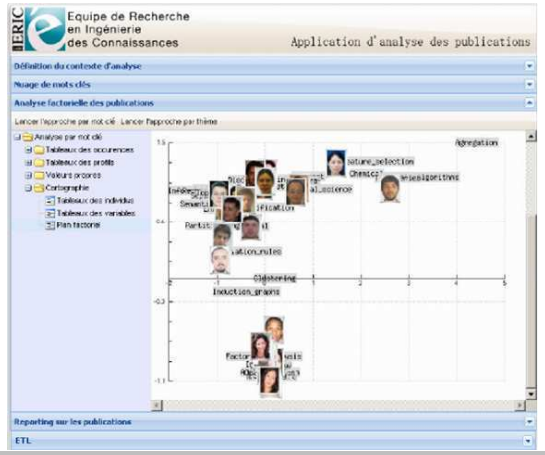
Choix de la partition
Critère de séparabilité des classes

Expérimentation, implémentation

- **Validation expérimentale**
  - Données classiques
  - Données complexes (mammographies, publications)

- **Plates-formes logicielles**
  - ***MiningCubes*** : analyse en ligne des données complexes
  - ***PUMA*** : application dédiée à l'analyse des publications





Conclusion scientifique

- **Problématique de l'analyse en ligne des données complexes**
- **Cinq verrous scientifiques abordés**
- **Premiers résultats intéressants et encourageants**
- **Démonstration de la pertinence et faisabilité** de combiner l'OLAP à d'autres techniques d'analyse
- **Evolution significative de l'OLAP**
  - S'adapter aux données complexes
  - Dépasser ses propres limites

Projet scientifique

- **Défi scientifique** : extraire et analyser (en ligne) la sémantique
- Vers une nouvelle génération d'analyse en ligne : **OLAP sémantique**
- **Création d'un nouveau thème de recherche** : problèmes théoriques, méthodologiques et technologiques
- **Verrous scientifiques**
  - Couvrir toutes les caractéristiques des données complexes
  - Modéliser toutes les formes de données complexes, leur sémantique et leurs liens
  - Analyser en ligne les données complexes
  - Intégrer les connaissances de l'utilisateur dans l'analyse
- **Formaliser l'OLAP sémantique**

Projet scientifique

- Projet interdisciplinaire entre les laboratoires ERIC et ICAR (Lyon 2-ENS-CNRS)
  - Interactions orales
  - Identification automatique de phénomènes complexes (conflit, plainte,...)
  - Base de données CLAPI : corpus oraux, transcriptions, documents XML
  - Entrepôt de corpus
  - Analyses appropriées
  - Prise en compte de la sémantique contenue dans les corpus

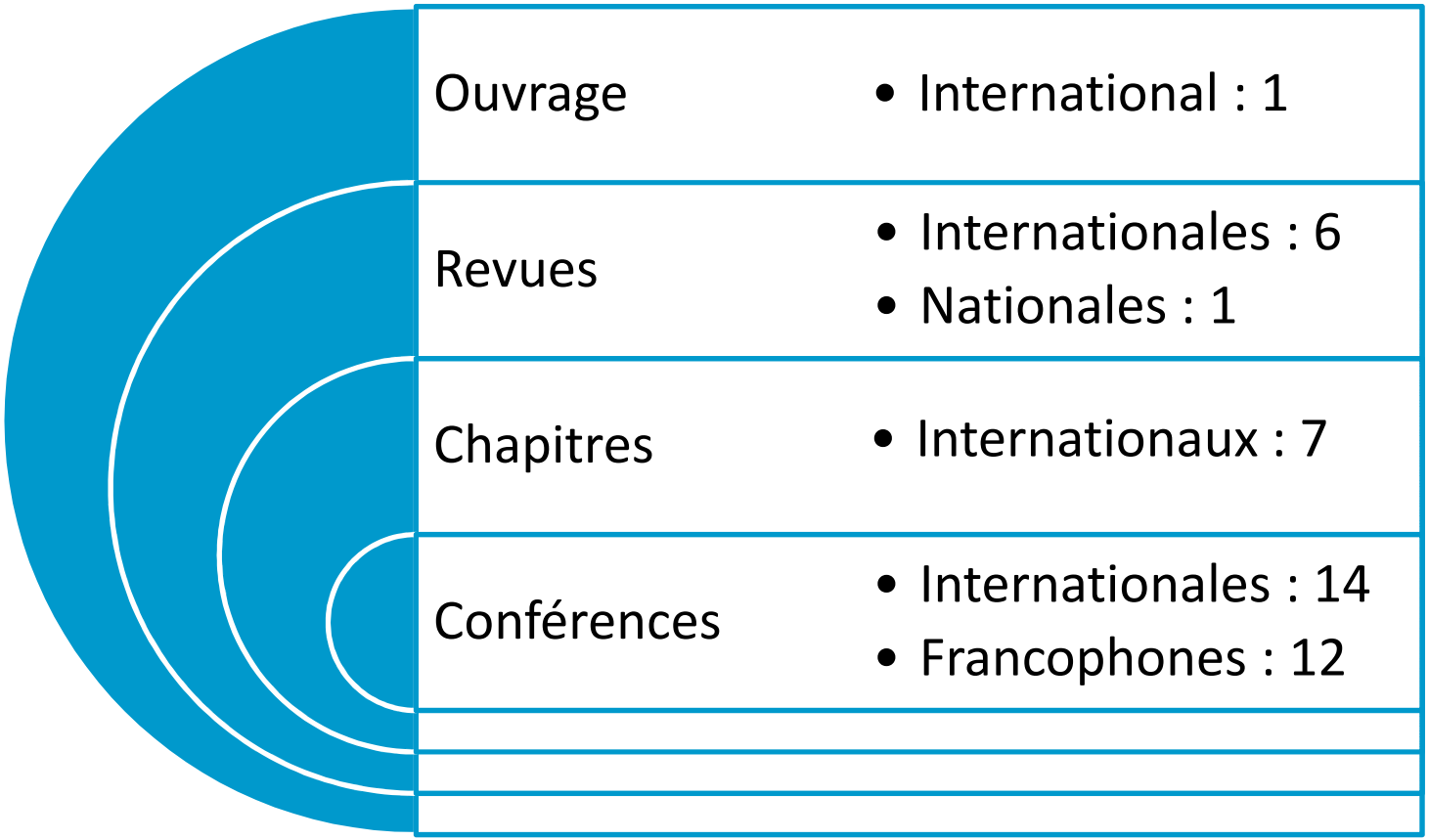
Encadrement scientifique

- **Co-encadrement de la thèse** de Riadh BEN MESSAOUD, 2003 - 2006

- **Participation à la thèse** d'Abdellah SAIR, Ecole Nationale des Sciences Appliquées, Agadir – Maroc, depuis septembre 2009

- **Formation à la recherche**
  - DEA, Riadh BEN MESSAOUD, 2003
  - Master recherche, Nourredine MOKTARI, 2005
  - Master recherche, Michel El RAHI, 2006
  - Master recherche, Slimane DJOUADI, 2006
  - Master recherche, Anouck BODIN-NIEMCZUK, 2007
  - Master recherche, Loic MABIT, 2009
  - Master recherche et professionnel, Youcef MECHEHOUD, Moussa ZOUBIRI, Caroline CHAILLET, 2010

Production scientifique



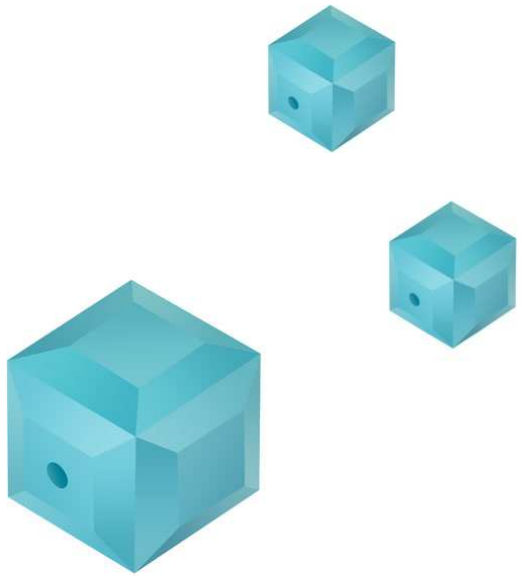
dont DMBI , IJWET, IJDWM, RTSI-ISI

dont CAISE, PKDD, DB&IS, DOLAP, CIKM, Inforsid, EGC

Animation et expertise scientifique

- **Comités éditoriaux ou de pilotage** : EDA, IJBET, WMCD
- **Comités de programme ou de lecture** : JDS 2003, PKDD2004, ISWC'04, ASD06 à ASD10, EDA06 à EDA11, IIS 2008, RNTI, TSI, …
- **Comités d'organisation** : SFC 1997, PKDD 2000, JDS 2003, EDA 2005, INFORSID 2013
- **Expertise** : dossiers de financement CIFRE-ANR
- **Groupes de travail ou associations scientifiques** : groupe de travail sur la Fouille de Données complexes , action Spécifique CNRS STIC GaFoDonnées : sous-groupe de travail GafOLAP), Société Française de Statistique (SFdS), Société Francophone de Classification (SFC)

Merci pour votre attention



LABO ERIC
EQUIPE DE RECHERCHE EN INGENIERIE DES CONNAISSANCES

UNIVERSITÉ LUMIÈRE LYON 2
UNIVERSITÉ DE LYON

UNIVERSITÉ DE LYON