Mission Interdisciplinaire Défi MASTODONS

Journée ICUBE, 7 novembre 2014

AMADOUER : Analyse de MAsses de DOnnées de l'Urbain et l'Environnement

A. Baskurt – S. Servigne – JF. Boulicaut

UMR 5600 UMR 5008 EA 4126 UM5 5205

Contexte

- MASTODONS
  - « Big Data »
  - « Data Science »

- AMADOUER
  - Absence de “commande” a priori
    - Masses de données hétérogènes et parfois volumineuses dans les domaines de l’urbain et de l’environnement
    - Questionnements variés
  - Travailler à l’émergence de communautés « data science »
  - Concepts – Généricité pour accompagner de bonnes ruptures méthodologiques en sciences

Singularités

- Contexte local

MI

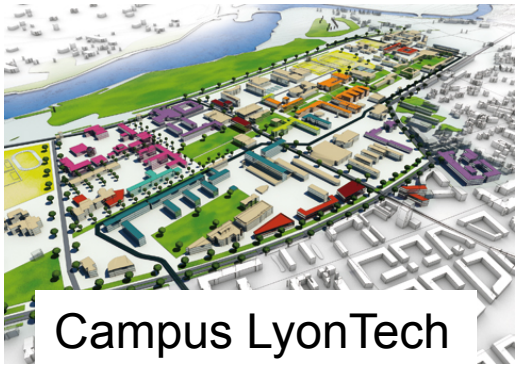
INEE
INSHS
INSIS
INS2I

cnrs

AMADOUER

GRANDLYON
communauté urbaine

IMU Intelligences des Mondes Urbains
UNIVERSITE DE LYON



Campus LyonTech

LIRIS

LGCIE
LABORATOIRE DE GÉNIE CIVIL & INGÉNIERIE ENVIRONNEMENTALE

EVS
Environnement Ville Société
cnrs
UMR 5600
Université de Lyon

Centre de Thermique de Lyon
CETHIL
UMR 5008

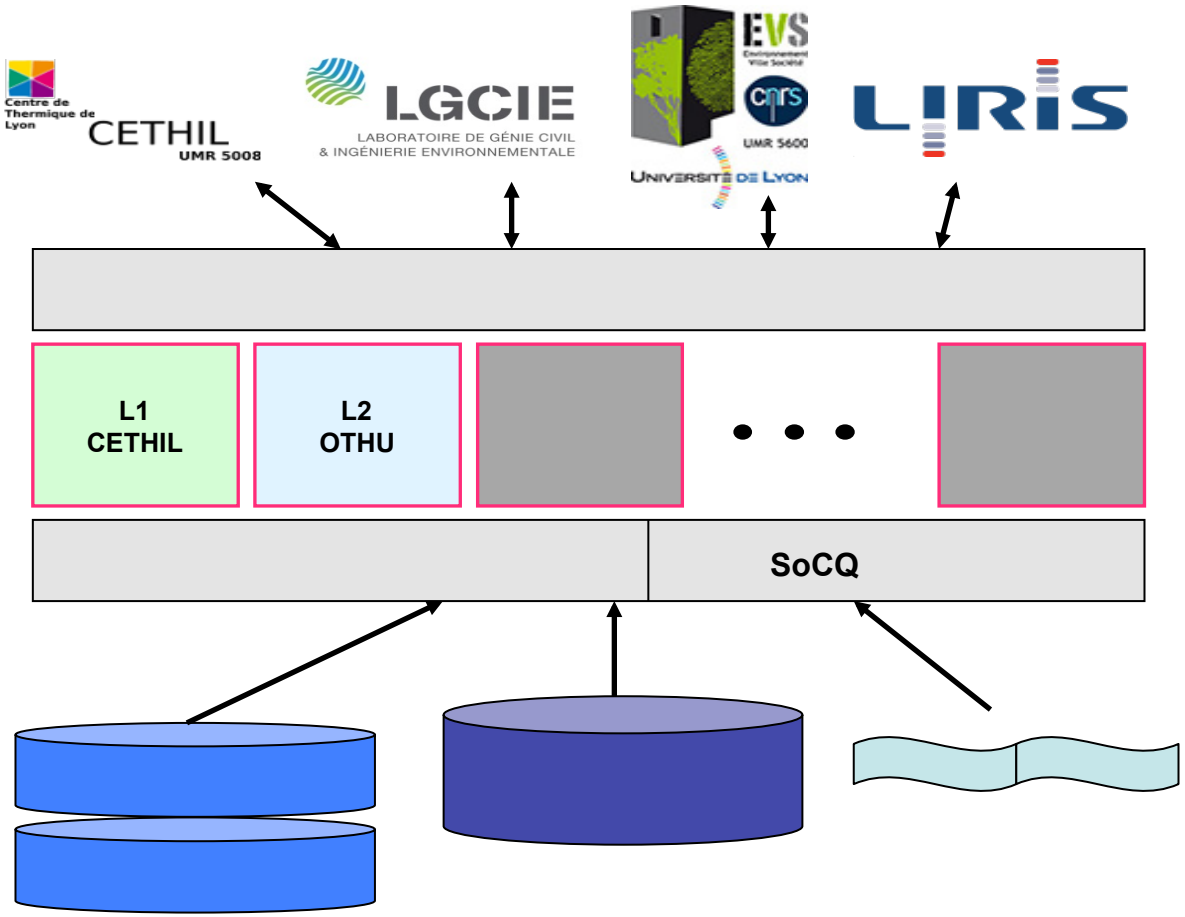
Bilan 2012

- Identification “Données/Questionnement”
  - EVS (INSHS, INEE)
    - Occulométrie dans un environnement équipé
  - CETHIL (INSIS)
    - Combustion
    - Changements de phase
    - Monitoring d’installation de production d’énergie solaire photovoltaïque
    - Monitoring énergétique de bâtiment intelligent
  - LGCIE-OTHU
    - Observatoire de Terrain en Hydrologie Urbaine
  - LIRIS
    - Plate-forme SoCQ4Home (mi-lourd CNRS)
  - GRAND LYON
    - Données : mobilité, énergie, cohésion sociale, etc

Flux vidéos

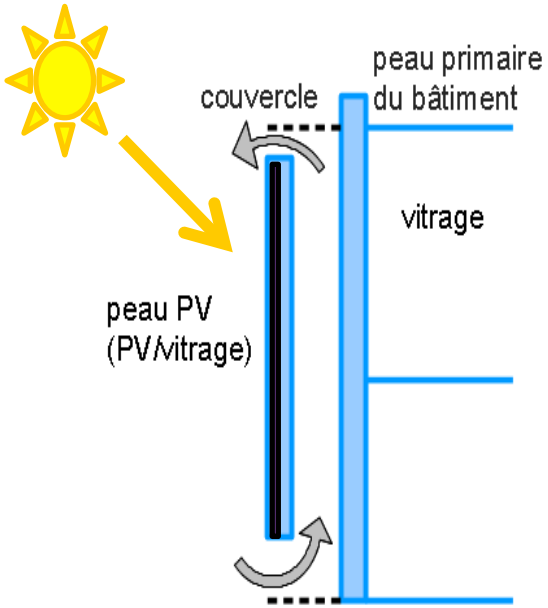
“Réseaux de capteurs”

Plate-forme


Centre de Thermique de Lyon
CETHIL
UMR 5008
LGCIE
LABORATOIRE DE GÉNIE CIVIL & INGÉNIERIE ENVIRONNEMENTALE
EVS
cnrs
UMR 5600
UNIVERSITÉ DE LYON
LIRIS
L1
CETHIL
L2
OTHU
SoCQ

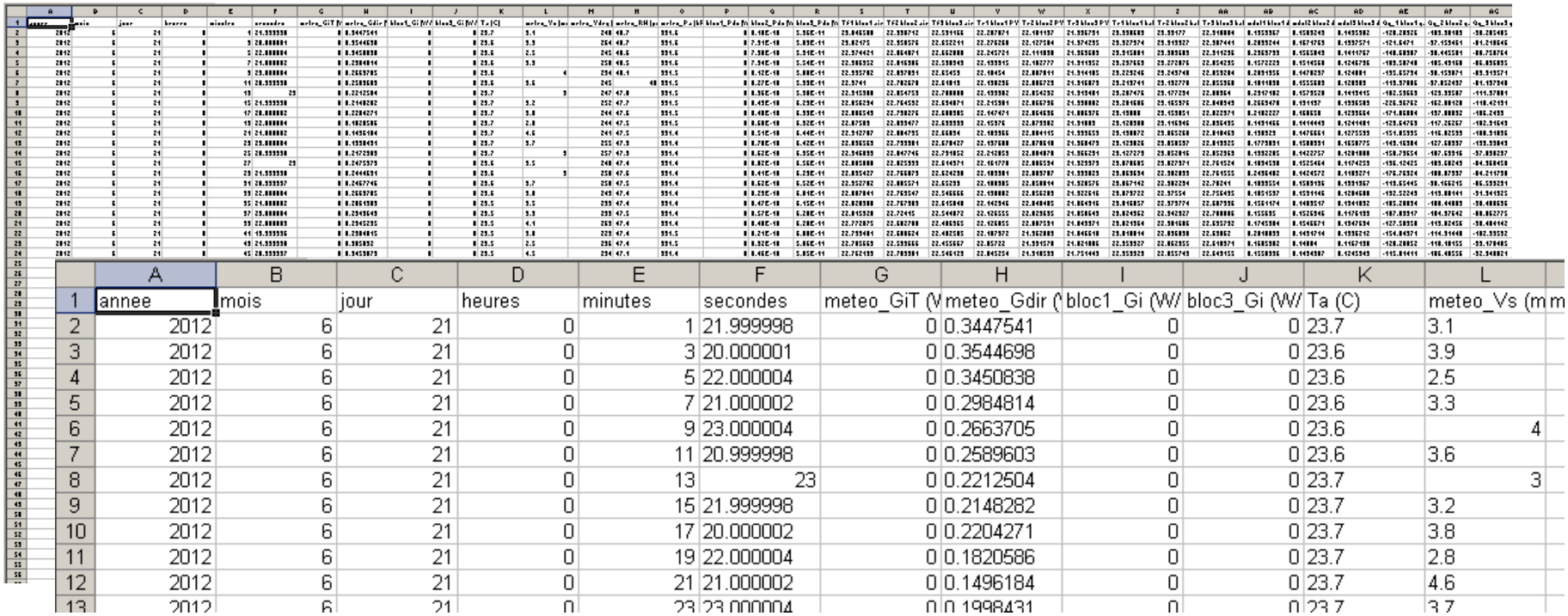
Données CETHIL (projet Ressources)

Réseaux de capteurs (électricité, anémomètres, pyranomètres, thermocouples, station météo+suntracker) ... Toutes les 2 minutes sur 1 an

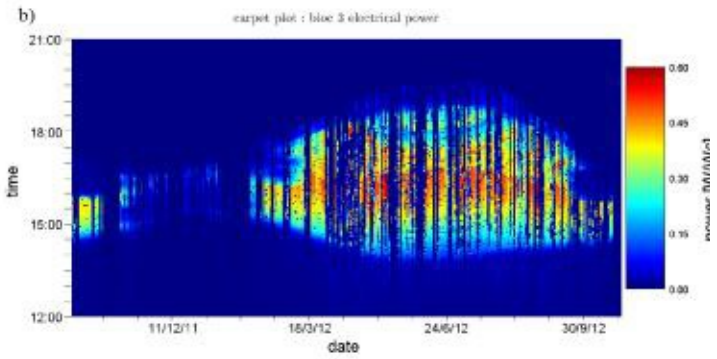


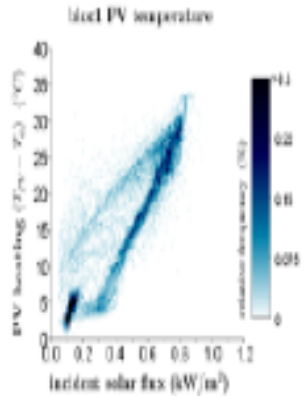
Questions prioritaires : Comportements “type” vs. “anormaux” (e.g., impact du vent), loi empirique pour lier le débit d’air aux autres variables

Données CETHIL (projet Ressources)



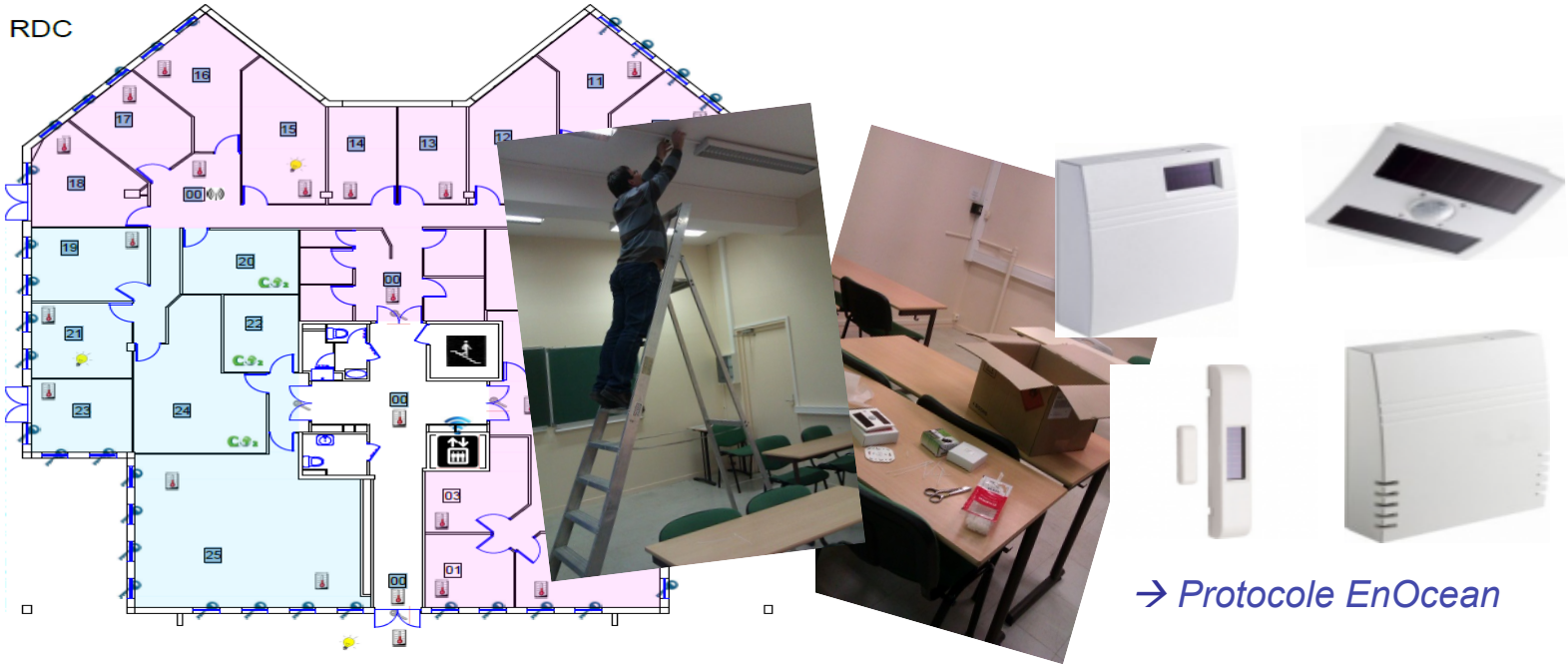
	A	B	C	D	E	F	G	H	I	J	K	L
1	annee	mois	jour	heures	minutes	secondes	meteo_GiT (V	meteo_Gdir ('	bloc1_Gi (W/	bloc3_Gi (W/	Ta (C)	meteo_Vs (m m
2	2012	6	21	0	1	21.999998	0	0.3447541	0	0	23.7	3.1
3	2012	6	21	0	3	20.000001	0	0.3544698	0	0	23.6	3.9
4	2012	6	21	0	5	22.000004	0	0.3450838	0	0	23.6	2.5
5	2012	6	21	0	7	21.000002	0	0.2984814	0	0	23.6	3.3
6	2012	6	21	0	9	23.000004	0	0.2663705	0	0	23.6	4
7	2012	6	21	0	11	20.999998	0	0.2589603	0	0	23.6	3.6
8	2012	6	21	0	13	23	0	0.2212504	0	0	23.7	3
9	2012	6	21	0	15	21.999998	0	0.2148282	0	0	23.7	3.2
10	2012	6	21	0	17	20.000002	0	0.2204271	0	0	23.7	3.8
11	2012	6	21	0	19	22.000004	0	0.1820586	0	0	23.7	2.8
12	2012	6	21	0	21	21.000002	0	0.1496184	0	0	23.7	4.6
13	2012	6	21	0	23	23.000004	0	0.1998431	0	0	23.7	3.7





Données : Bâtiment intelligent

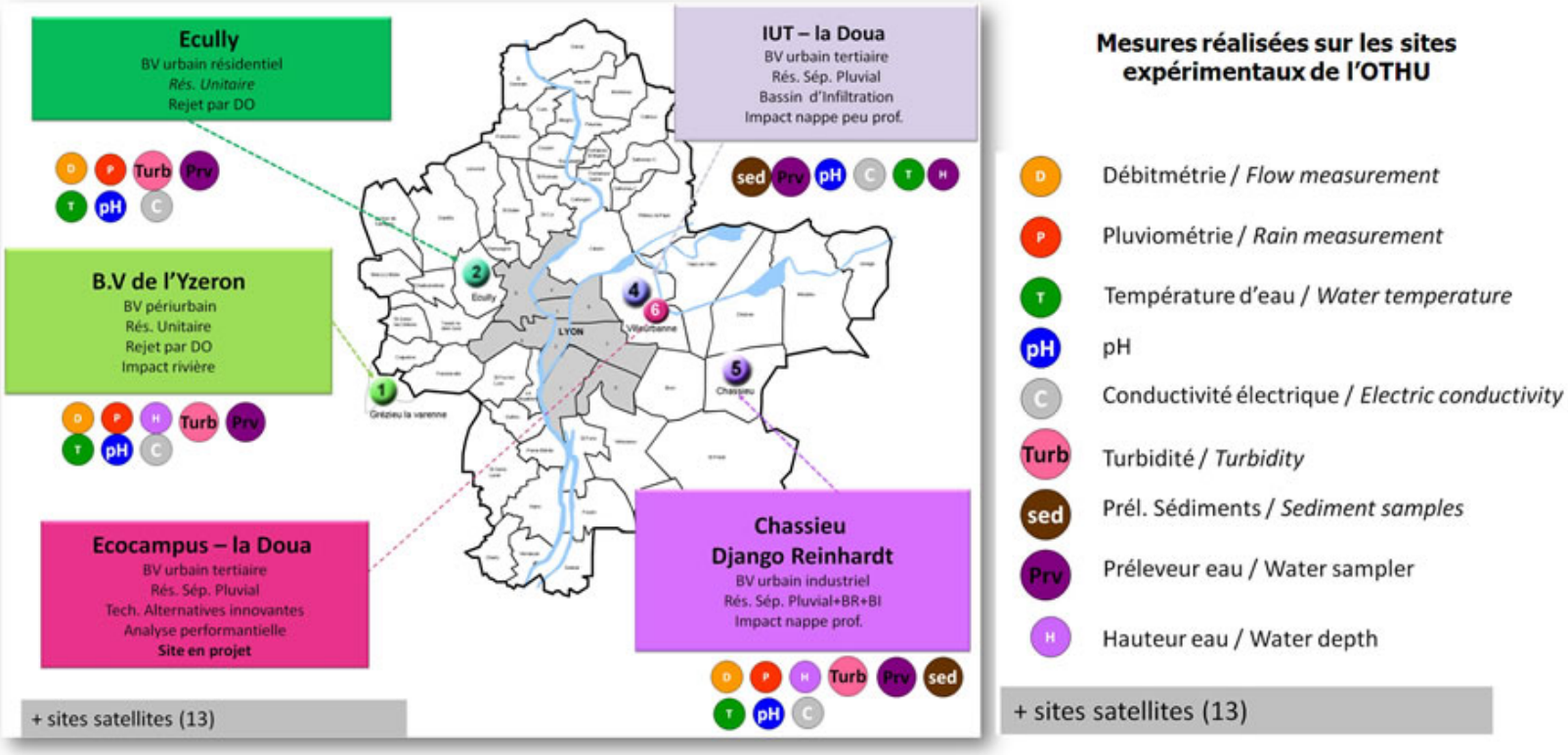
- Plateforme MARBRE (INSA : LIRIS-CETHIL-EVS)
  - Réseau 300 capteurs sans-fils (température, humidité, Co2, luminosité, ouverture porte/fenêtre)
  - Emission capteurs : toutes les 15 minutes ou selon événement
  - Enquêtes usagers



Question prioritaire : Validation de modèles physiques avec prise en compte des usages

Données OTHU

Réseaux de capteurs opérationnels depuis 2001 (Observatoire) : toutes les 2 minutes 7j/7 24h/24



Questions prioritaires : qualité des données, "la constitution des flux polluants n'est pas comprise" (multi-échelles, nouvelles variables, ...)

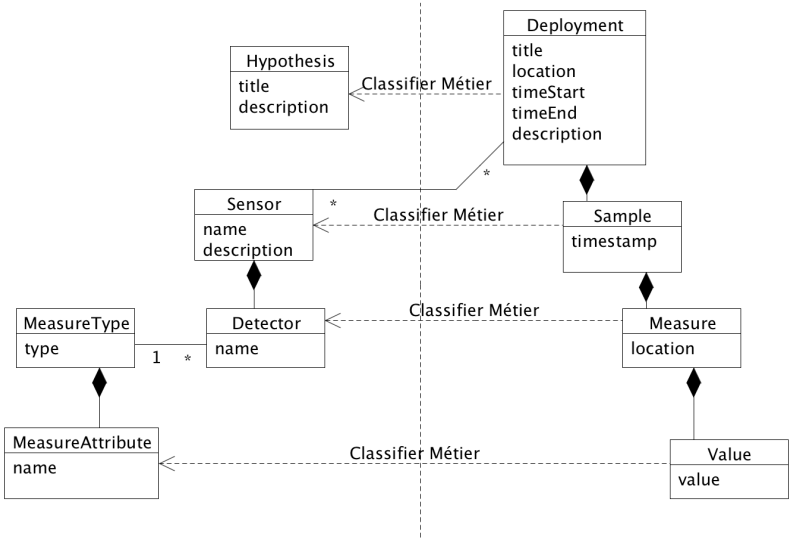
Verrous traités depuis 2013

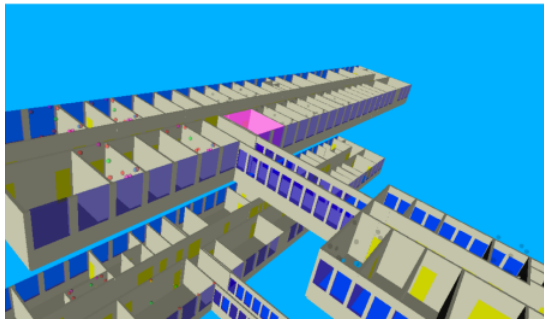
- Modèles de données génériques adaptés aux flux
- Concepts et langage d'exploration données/flux
- Découverte de connaissances (fouille/modèles) : construction de (nouveaux) domaines de motifs
- Instanciation en thermique
  - Validation de modèles physiques
  - Détection d'anomalies
  - Interactions et corrélations indirectes
  - Conception de nouveaux modèles dynamiques (modélisations non-physiques, comportementales, etc.) pour une meilleure prédiction de la production instantanée ou cumulée (injection réseau, couplage bâtiment)

Résultats (1)

Modèles de données

- Définition et formalisation du modèle Virtual Generic Sensor (VGS) collaboration LIRIS (INS2I) et EVS (INSHS-INEE)
- Application à la prise en compte des usages en thermique du bâtiment.





Visualisation Web 3D

Sensor1	HumidityDetector1	10/02/2013 20:05:00	Humidity	Humidity	1000	(2,2,3)@502:309	42
Sensor1	HumidityDetector1	10/02/2013 20:05:01	Humidity	Humidity	1001	(2,2,3)@502:309	43
Sensor1	HumidityDetector1	10/02/2013 20:05:02	Humidity	Humidity	1002	(2,2,3)@502:309	44
Sensor1	HumidityDetector1	10/02/2013 20:05:03	Humidity	Humidity	1003	(2,2,3)@502:309	42
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
Sensor3	Question1	10/02/2013 20:05:00	Question	Humor	3000	(1,1,3)@502:411	5
Sensor3	Question1	10/02/2013 20:05:00	Question	Vitality	3001	(1,1,3)@502:411	7
Sensor3	Question1	10/02/2013 20:05:05	Question	Humor	3002	(1,1,3)@502:410	2
Sensor3	Question1	10/02/2013 20:05:05	Question	Vitality	3003	(1,1,3)@502:410	2

Résultats (2)

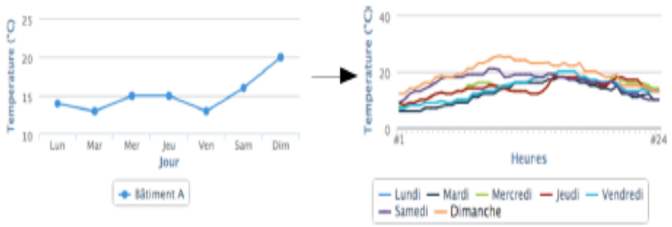
Exploration → Astéroïdes et Langage

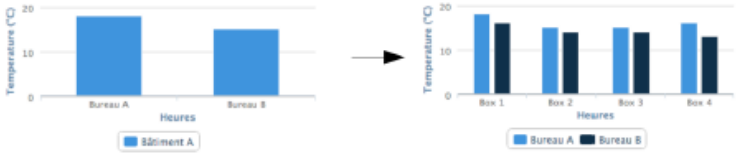
- Définition et formalisation du concept d'Astéroïde
- Exploration multidimensionnelle interactive par agrégations dynamiques multi-échelles de données, pilotée par l'utilisateur

Schéma d'Astéroïde sur un schéma de Série Multidimensionnelle: sous-ensemble de dimensions, niveau "courant" de chaque dimension et schéma des valeurs

- Soit un Schéma de Série Multidimensionnelle $\mathcal{S} = (\mathcal{W}_\mathcal{S}, \mathcal{V}_\mathcal{S})$
- Schéma d'Astéroïde $\mathcal{X} = (\mathcal{W}_\mathcal{X}, [[l_j]]^j_{\mathcal{W}_\mathcal{X}}, \mathcal{V})$ sur $\mathcal{S}$
  - $\mathcal{W}_\mathcal{X} = \langle \mathcal{D}_{j_1}, \ldots, \mathcal{D}_{j_m} \rangle \subseteq \mathcal{W}_\mathcal{S}$ liste d'un sous-ensemble des dimensions du schéma de Série Multidimensionnelle $\mathcal{W}_\mathcal{S} = \langle \mathcal{D}_1, \ldots, \mathcal{D}_n \rangle$
  - $l_j \in L_{\mathcal{D}_j}$ niveau de la dimension $\mathcal{D}_j \in \mathcal{W}_\mathcal{X}$
  - $\mathcal{V}$ symbole de "relation" pour les valeurs des points
    - $schema(\mathcal{V}) \subset \mathcal{A}$

Note: un schéma de Série Multidimensionnelle $\mathcal{S} = (\mathcal{W}_\mathcal{S}, \mathcal{V}_\mathcal{S})$ se "transpose" en schéma d'Astéroïde $\mathcal{X} = (\mathcal{W}_\mathcal{S}, [[l_i^\bullet]]^i_{\mathcal{W}_\mathcal{X}}, \mathcal{V}_\mathcal{S})$ avec $l_i^\bullet$ le niveau de base de la dimension $\mathcal{D}_i \in \mathcal{W}_\mathcal{S}$.





Project (i.e., select a subset of values and/or compute new values)

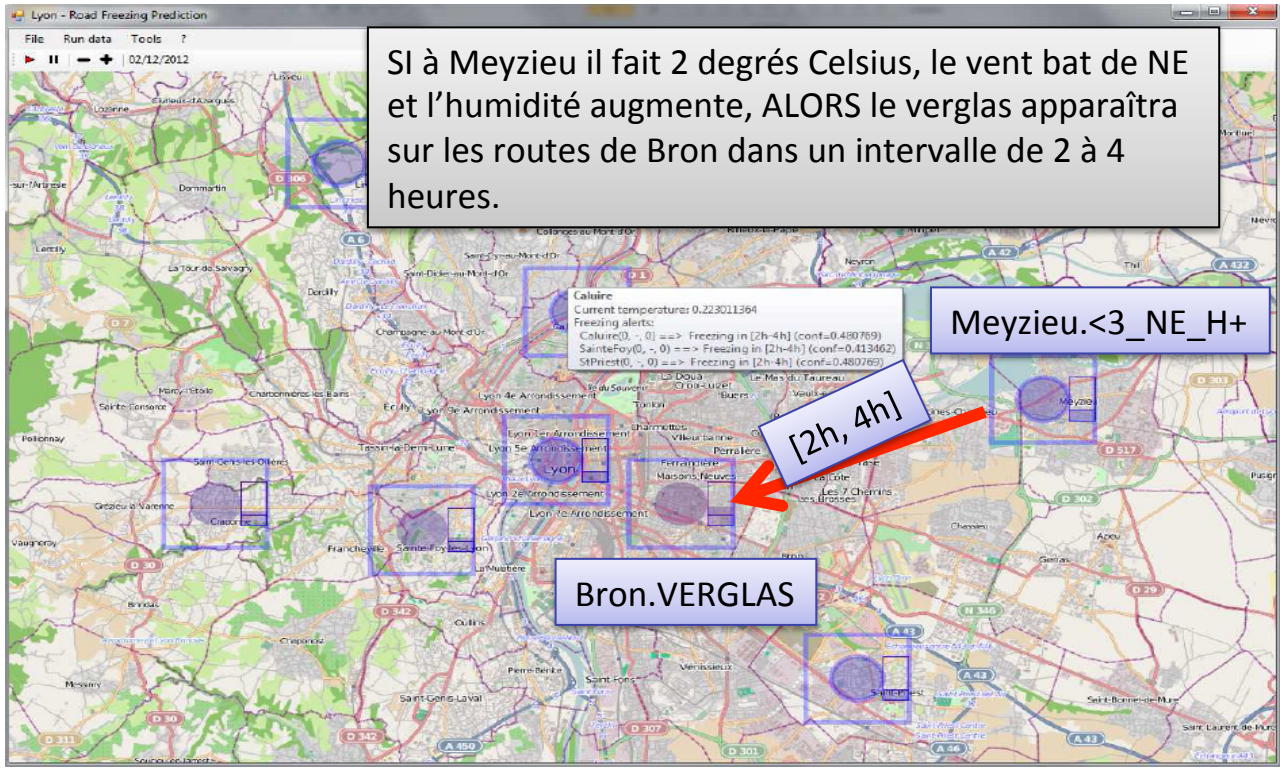
Calcul de nouvelles valeurs par l'application d'une fonction sur les valeurs de chaque point.

- Soit $x = ([[d_i]]^i_{\mathcal{W}_\mathcal{X}}, [[M_i]]^i_{\mathcal{W}_\mathcal{X}}, V)$ sur $\mathcal{X} = (\mathcal{W}_\mathcal{X}, [[l_i]]^i_{\mathcal{W}_\mathcal{X}}, \mathcal{V}_\mathcal{X})$
- Soit une fonction (simple) $\mathcal{F}^\circ_s : \mathcal{V}_\mathcal{X} \to \mathcal{V}'$
- Résultat $x' = ([[d_i]]^i_{\mathcal{W}_\mathcal{X}}, [[M_i]]^i_{\mathcal{W}_\mathcal{X}}, V')$ sur $\mathcal{X}' = (\mathcal{W}_\mathcal{X}, [[l_i]]^i_{\mathcal{W}_\mathcal{X}}, \mathcal{V}')$
  - $V'(p) = \mathcal{F}^\circ_s(V(p))$
- Notation $x' = \Upsilon_{\mathcal{F}^\circ_s}(x)$ *(lettre υ (upsilon) majuscule)*

Résultats (3)

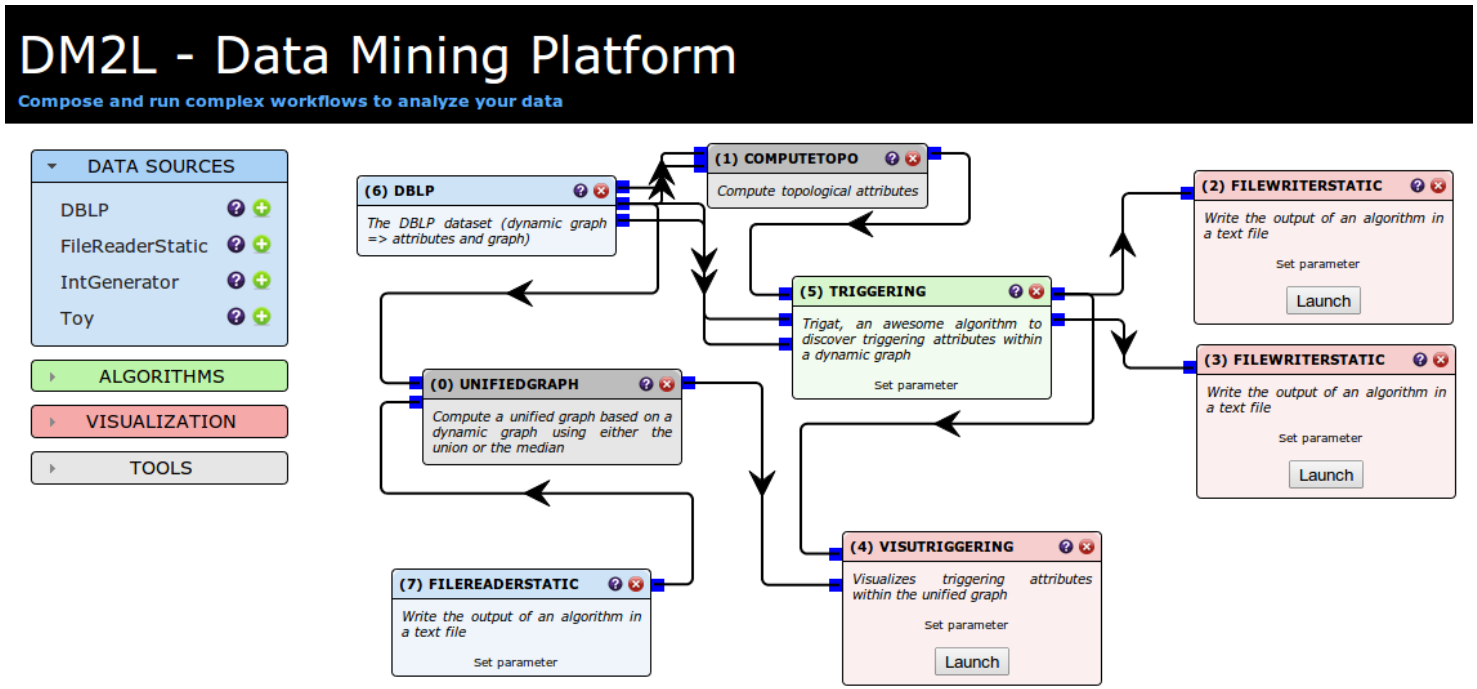
Extraction de connaissances → KDD

- Découverte de dépendances temporelles dans l'analyse de flux de données du Grand Lyon : application au salage de la voirie



Résultats (4)

Plateforme AMADOUER



A noter

- C'est dans la généricité des concepts proposés et implémentés que réside la valeur ajoutée
  - Langage de workflow pour une capitalisation des développements de prototypes en exploration et fouille des données
  - Langages et modèles adaptés aux données issues de réseaux de capteurs … en évolution
  - Algorithmes de fouille de données génériques (e.g., analyse des dépendances temporelles entre flux de données)

Travail en cours sur 2014

- Développement de démonstrateurs sur la plateforme AMADOUER
- Synthèse sur nos analyses de masses de données urbaines produites par le Grand Lyon (bientôt métropole)
- Rapprochement avec ANIMITEX et travail sur une feuille de route concernant les verrous « Hétérogénéité & Big Data » (illustrations dans les sciences de l'environnement)
- Coopération avec le PPME de l'Université de Nouvelle Calédonie (Nazha Selmaoui, Frédéric Flouvat)
  - Exploitation de modèles experts pour la dérivation de contraintes

$$\{\tau \in L \mid q(\tau, r) \text{ est vrai}\}$$

Conclusion (1)

- Des expériences en cours sur l'émergence de "petites" communautés "Data Science"
  - Thermique, Energie, Environnement
  - Sciences du Vivant, Physique, Mathématique, Sciences Humaines et Sociales, etc
- A confronter à d'autres projets
  - Notamment dans le contexte MASTODONS
  - Actions avec ANIMITEX (Juin 2014, Novembre 2014)
- Tâche ingrate mais nécessaire ;-)

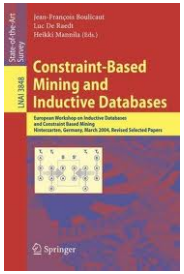
Conclusion (2)

- Il faut accompagner des ruptures méthodologiques de recherche en identifiant des solutions génériques
  - A l'échelle de, e.g., MASTODONS, il y a donc des REX dans des contextes très différents
    - La valorisation de masses de données identifiées (par exemple PETASKY) avec des défis très "Volume"
      - E.g., passage à l'échelle de tel ou tel algorithme sur les images LSST
    - L'identification des difficultés et la proposition d'éléments de méthodes pour réussir des expériences en science des données.
      - E.g., méthodologie de co-construction et la validation de domaines de motifs

Co-construction de domaines de motifs

- Le propriétaire des données au coeur des processus de création de la valeur ajoutée

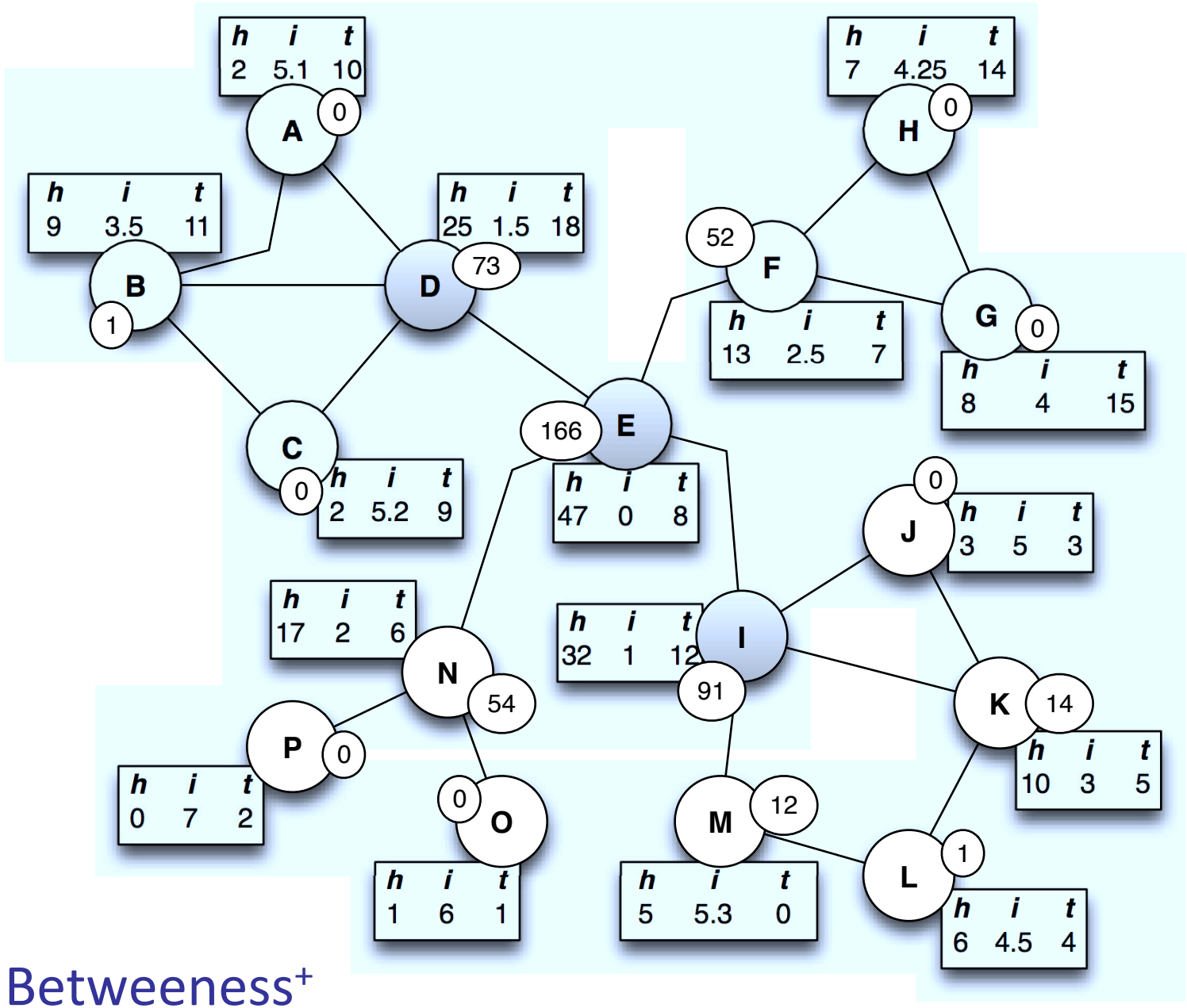
  (L,C,M,r) Langage de motifs
  Contraintes Primitives
  Modèle des connaissances a priori
  Données



  Requêtes inductives $\{\tau \in L \mid q(\tau, M, r) \text{ est vrai}\}$

- Comment concevoir et prototyper un domaine de motif ?

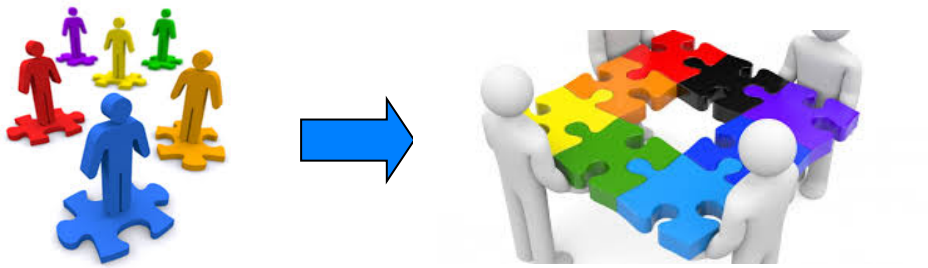
  Déclaratif vs. Calcul, Générique vs. Ad-hoc, Exact vs. Approché, Préférences et souplesse, etc



$h^+ i^-$ Betweeness$^+$

Conclusion (3)

- Valorisation de notre REX sur l'émergence de communautés en science des données
  - Impact sur les formations ?
  - GDR ?



Participants AMADOUER

M. Cottet, J-M. Deleuil, J-Y. Toussaint, S. Vareilles

J. Bonjour, L. Gaillard, S. Giroux, J. Jay, C. Ménézo, H. Pabiou

A. Baskurt, J-F. Boulicaut, G. Brochet, S. Fenet, Y. Gripay, M. Keytoue, M. Plantevit, Y. Pitarch, M. Scuturici, C. Robardet, S. Servigne, C. Wolf (PPME N. Selmaoui, F. Flouvat – Université de Nouvelle Calédonie)

S. Barraud, N. Walcker, L. Bacot