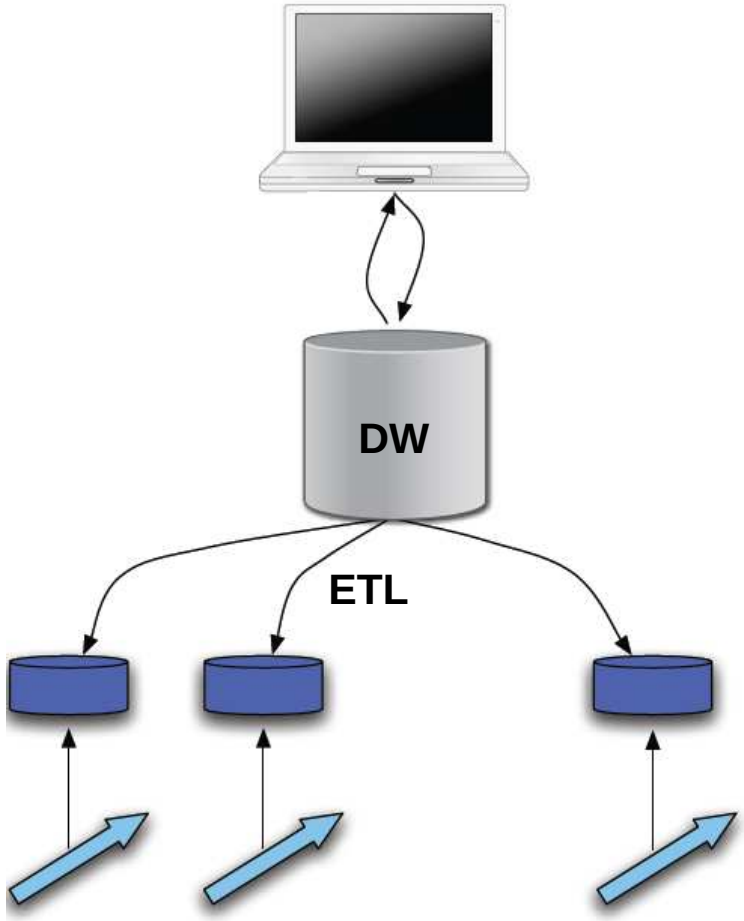
Echantillonnage optimisé de données temporelles pour l'alimentation des entrepôts de données

Raja Chiky et Georges Hébrail
ENST-Paris
EDF R&D
07 juin 2007

Introduction (1/3)

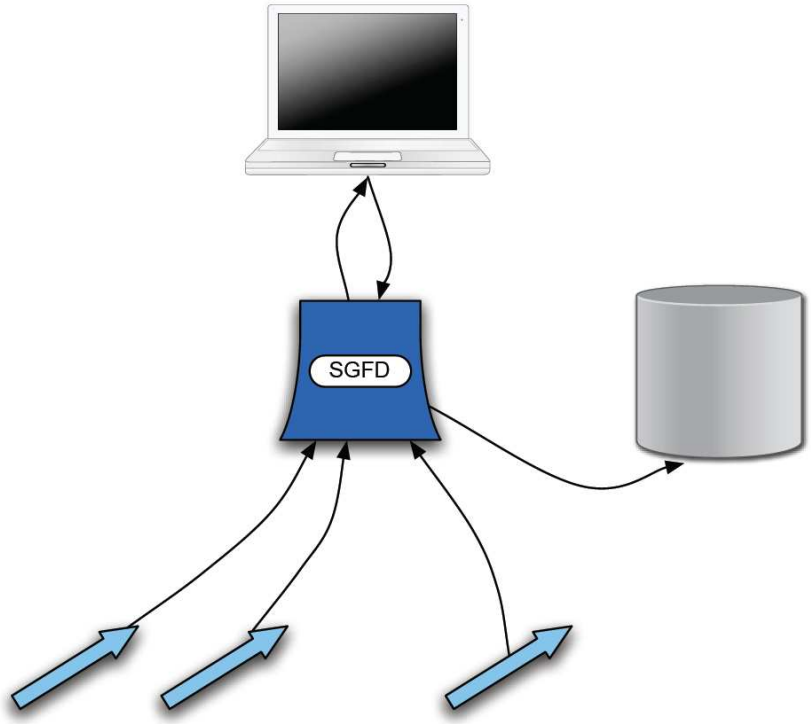
Contexte

- Plusieurs applications génèrent des données sous forme de flux à partir de capteurs ou sources de données à haut débit
- Nécessité de prendre en compte la volumétrie et la vitesse d'arrivée des données provenant des capteurs pour l'historisation dans les entrepôts de données
- Les méthodes ETL ne répondent pas à des besoins de traitement au fil de l'eau



Introduction (2/3)
SGFD

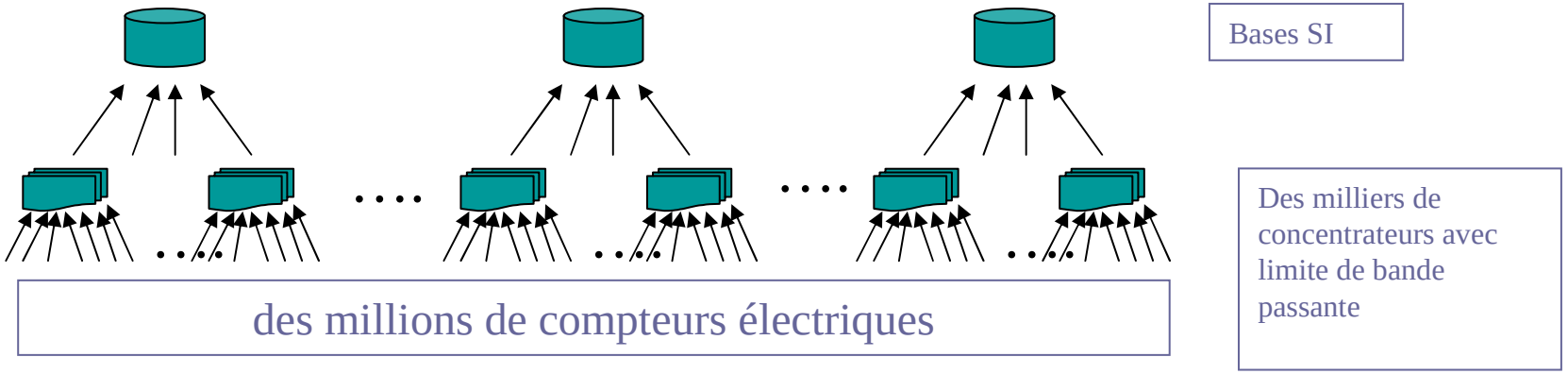
- Objectifs:
  - Traiter les flux au fil de l'eau sans stocker les données
  - Répondre à des requêtes continues (alarmes)
  - Éviter la surcharge du système
  - …
- Quelques SGFD existants:
  - Projets de recherche: STREAM, TelegraphCQ, Aurora
  - Commerciaux: StreamBase, Aminsight



Introduction (3/3)

Exemple

- Exemple d'architecture de récupération de données à partir de compteurs électriques communicants
  - Déploiement massif de compteurs communicants : consommations d'énergie électrique relevées à un pas très fin (à partir de la seconde) => courbes de charges CDC

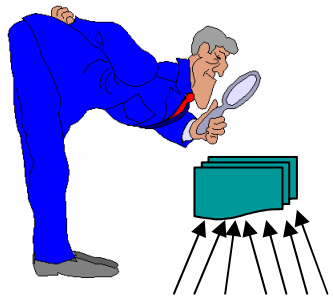


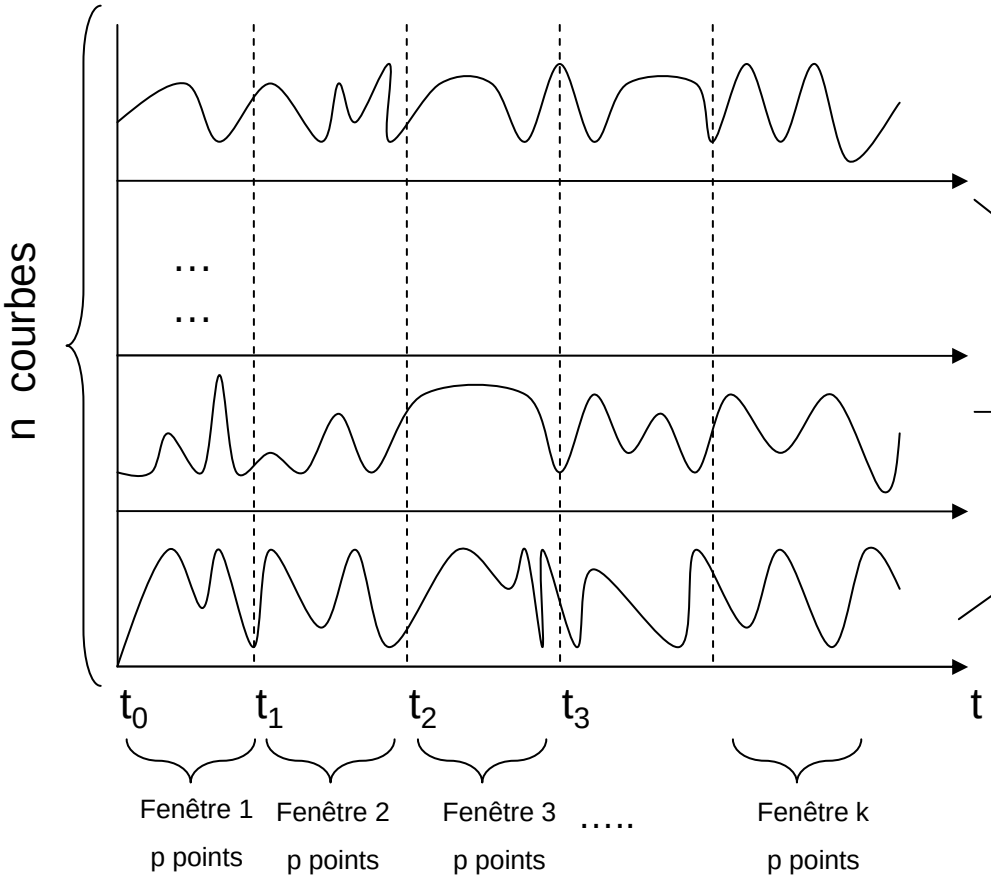
- Objectif: alimenter les entrepôts de données à partir de « flux de données » distribués

Plan de l'exposé

- Formulation du problème
- Méthode de résolution
- Expérimentations
- Traitement global du flux
- Travaux en cours

Formulation du problème (1/3)



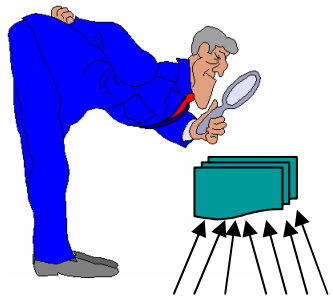


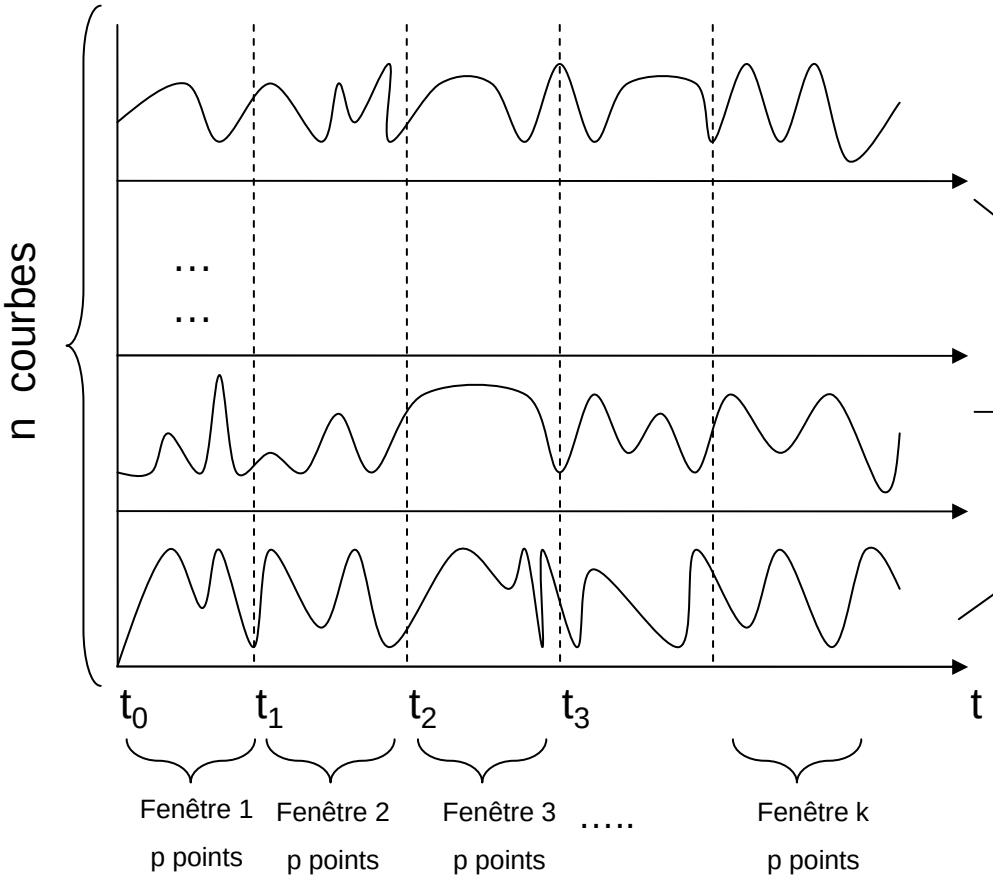
Concentrateur:

s: #points communicables sur une fenêtre t (s< n*p)

m: #points minimum à prélever par courbe sur une fenêtre t

Formulation du problème (1/3)





Concentrateur:

s: #points communicables sur une fenêtre t (s< n*p)

m: #points minimum à prélever par courbe sur une fenêtre t

Trouver la meilleure « politique d'échantillonnage » pendant une fenêtre t

Formulation du problème (2/3)

-Echantillonnage régulier-

- Chaque courbe i est échantillonnée à un pas j_i
  - j_i: saut à effectuer entre deux données à sélectionner
    - j_i=1: tous les points sont sélectionnés
    - j_i=2: un point sur deux est sélectionné
    - …
  - $f_i=1/j_i$: fréquence (ou taux) d'échantillonnage

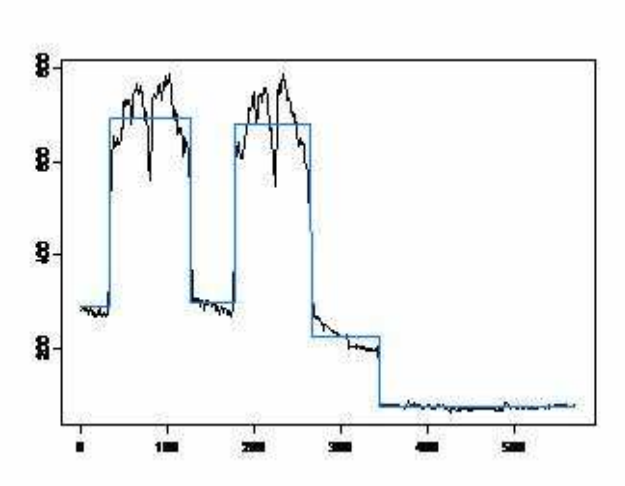
- Trouver les pas d'échantillonnage j_i tels que l'erreur d'échantillonnage est minimum et

$$\forall i \in 1...n \; j_i \leq \left\lfloor \frac{p}{m} \right\rfloor \quad \& \quad \sum_{i=1}^{n} \left\lfloor \frac{p}{j_i} \right\rfloor \leq s$$

Formulation du problème (3/3)

-Echantillonnage irrégulier-

- Segmentation de la courbe en m épisodes
  - Approximation par une fonction en escalier
  - Les épisodes peuvent être de durées différentes => « irrégulier »



- Trouver l'ensemble des segments et les moyennes associées de façon à minimiser l'erreur d'échantillonnage et en respectant les contraintes précédentes

Plan de l'exposé

- Formulation du problème
- Méthode de résolution
- Expérimentations
- Traitement global du flux
- Travaux en cours

Méthode de résolution (1/2)

- Erreur d'échantillonnage d'une courbe: Somme des erreurs quadratiques SSE

$$SSE(C,\hat{C}) = \sum_{i=1}^{p}(c_i - \hat{c}_i)^2$$

- Notations

$$m' = \left\lfloor \frac{p}{m} \right\rfloor$$

$W_{n*m'}$: matrice des erreurs de n lignes et m' colonnes.

$w_{ij} \in W_{n*m'}$ est la somme des erreurs quadratiques SSE obtenue si on:

- applique un pas d'échantillonnage j à la courbe i
- découpe de façon optimale la courbe i en $\left\lfloor \frac{p}{j} \right\rfloor$ segments

- Résoudre:

$$\text{Minimiser} \sum_{i=1}^{n}\sum_{j=1}^{m'}(W_{ij} \times X_{ij})$$

sous les contraintes :

$$\begin{cases} X_{ij} = 0 \text{ ou } 1 & \\ \sum_{j=1}^{m'} X_{ij} = 1 & \text{i de 1 à n} \\ \sum_{i=1}^{n}\sum_{j=1}^{m'}(\left\lfloor \frac{p}{j} \right\rfloor \times X_{ij}) \leq s & \text{i de 1 à n} \end{cases}$$

Méthode de résolution (2/2)

Résoudre:

$$\text{Minimiser} \sum_{i=1}^{n} \sum_{j=1}^{m'} (W_{ij} \times X_{ij})$$

sous les contraintes :

$$\begin{cases} X_{ij} = 0 \text{ ou } 1 & \text{(1)} \\ \sum_{j=1}^{m'} X_{ij} = 1 & \text{(2)} \quad \text{i de 1 à n} \\ \sum_{i=1}^{n} \sum_{j=1}^{m'} \left(\left\lfloor \frac{p}{j} \right\rfloor \times X_{ij} \right) \leq s & \text{(3)} \quad \text{i de 1 à n} \end{cases}$$

(1) $X_{ij}=1$ => le pas d'échantillonnage j (ou $\left\lfloor \frac{p}{j} \right\rfloor$ segments)est affecté à la courbe i
(2) => une seule valeur de j par courbe
(3) => la somme des données prélevées ne doit pas dépasser le seuil s

=> Problème d'optimisation linéaire

Méthode de résolution (2/2)

Résoudre:

$$\text{Minimiser} \sum_{i=1}^{n} \sum_{j=1}^{m'} (W_{ij} \times X_{ij})$$

sous les contraintes :

$$\begin{cases} X_{ij} = 0 \text{ ou } 1 & \\ \sum_{j=1}^{m'} X_{ij} = 1 & \text{i de 1 à n} \\ \sum_{i=1}^{n} \sum_{j=1}^{m'} (\left\lfloor \frac{p}{j} \right\rfloor \times X_{ij}) \leq s & \text{i de 1 à n} \end{cases} \quad \begin{matrix} (1) \\ (2) \\ (3) \end{matrix}$$

W_{ij}	[,1]	[,2]	[,3]	[,4]
[1,]	0	4.0	10.0	13.3
[2,]	0	4.3	8.1	12.4
[3,]	0	4.6	8.0	16.1
[4,]	0	4.1	7.0	13.1

→

	[,1]	[,2]	[,3]	[,4]
[1,]	0	1	0	0
[2,]	0	0	0	1
[3,]	0	0	1	0
[4,]	0	0	1	0

(1) X_{ij}=1 => le pas d'échantillonnage j (ou $\left\lfloor \frac{p}{j} \right\rfloor$ segments)est affecté à la courbe i
(2) => une seule valeur de j par courbe
(3) => la somme des données prélevées ne doit pas dépasser le seuil s

=> Problème d'optimisation linéaire

Plan de l'exposé

- Formulation du problème
- Méthode de résolution
- Expérimentations
- Traitement global du flux
- Travaux en cours

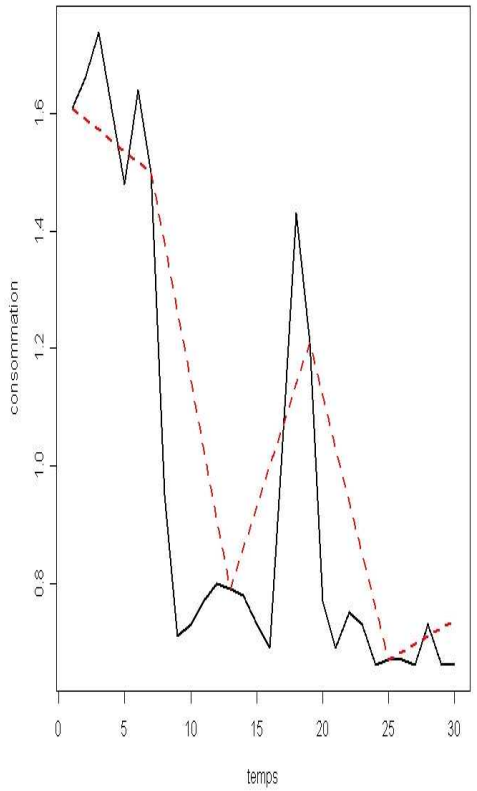
Expérimentations (1/4)

- Méthodes d'interpolation
  - Pour l'échantillonnage régulier
    - Interpolation linéaire

$$\hat{c}_i = \begin{cases} c_i & \text{si } i \text{ } modulo \text{ } j = 1 \\ f(i) & \text{sinon} \end{cases}$$

a<i<b, a et b deux points sélectionnés

$$f(i) = \frac{i-b}{a-b}c_a - \frac{i-a}{a-b}c_b$$



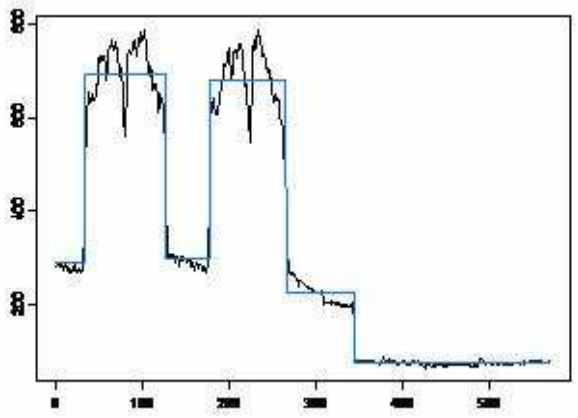
Expérimentations (1/4)

- Méthodes d'interpolation
  - Pour l'échantillonnage régulier
    - Interpolation linéaire

$$\hat{c}_i = \begin{cases} c_i & \text{si } i \text{ } modulo \text{ } j = 1 \\ f(i) & \text{sinon} \end{cases}$$

a<i<b, a et b deux points sélectionnés

$$f(i) = \frac{i-b}{a-b}c_a - \frac{i-a}{a-b}c_b$$



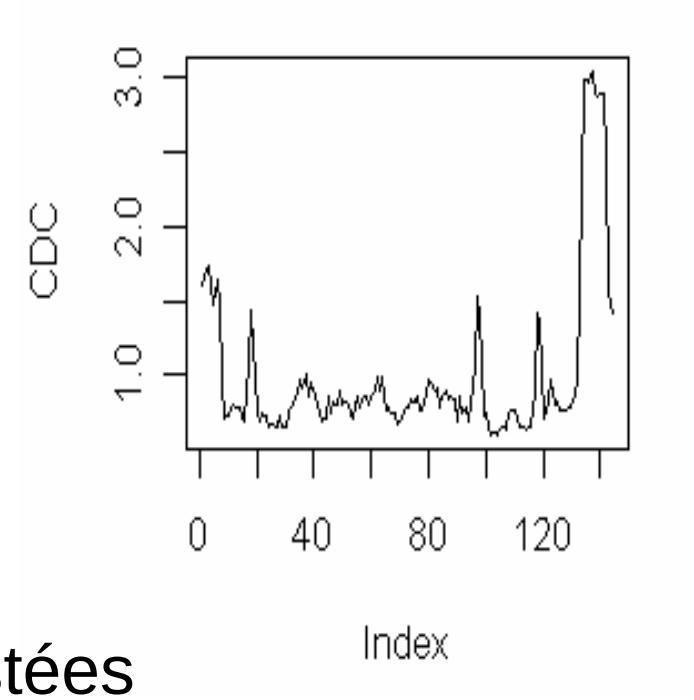
  - Pour l'échantillonnage irrégulier
    - Construction de la courbe échantillonnée en escalier à partir du nombre de points constituant les épisodes et les moyennes associées => $2*\left\lfloor \frac{p}{j} \right\rfloor$ données envoyées au concentrateur

Expérimentations (2/4)

- Pour les expérimentations:
  - 1000 courbes de charge

  CDC de 144 points, m=7

$$(m' = \left\lfloor \frac{144}{7} \right\rfloor = 20)$$



- Plusieurs valeurs de seuil s testées

Expérimentations (3/4)

seuil	144000	72000	48000	28800	20571	14400	9600
SSE opt reg	0	13.88	47.88	154.51	330.83	793.29	2336.9
SSE opt irreg	0	13.29	42.86	132.16	271.01	607.5	1730.6
SSE équirép. reg	0	1956	2821	4245	5501	7024	9087
SSE équirép. irreg	0	505.8	860.8	1535	2042.6	2761.5	4677.1

TAB. 1 – *tableau récapitulatif. SSE : Somme de l'écart quadratique. opt : pas d'échantillonnage (nb segments) obtenus par optimisation, et équirep. : même pas d'échantillonnage (nb segments) pour les CDCs. reg : échantillonnage régulier. irreg : échantillonnage par segmentation.*

Expérimentations (3/4)

seuil	144000	72000	48000	28800	20571	14400	9600
SSE opt reg	0	13.88	47.88	154.51	330.83	793.29	2336.9
SSE opt irreg	0	13.29	42.86	132.16	271.01	607.5	1730.6
SSE équirép. reg	0	1956	2821	4245	5501	7024	9087
SSE équirép. irreg	0	505.8	860.8	1535	2042.6	2761.5	4677.1

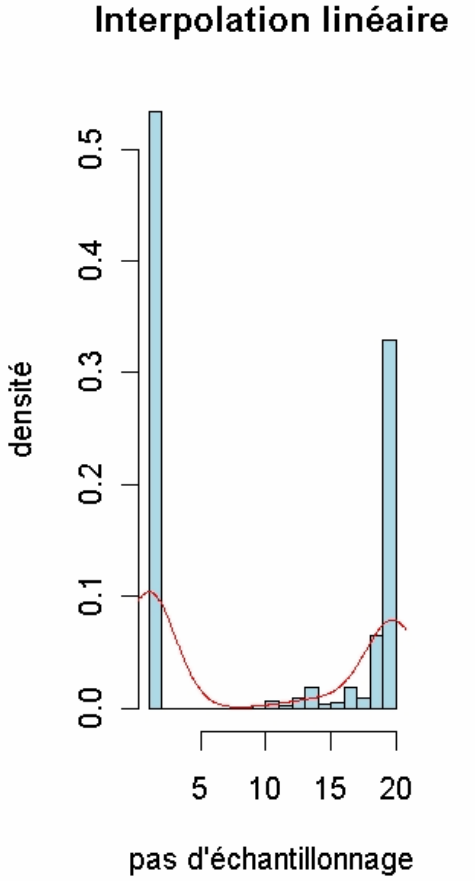
TAB. 1 – *tableau récapitulatif. SSE : Somme de l'écart quadratique. opt : pas d'échantillonnage (nb segments) obtenus par optimisation, et équirep. : même pas d'échantillonnage (nb segments) pour les CDCs. reg : échantillonnage régulier. irreg : échantillonnage par segmentation.*

Expérimentations (3/4)

seuil	144000	72000	48000	28800	20571	14400	9600
SSE opt reg	0	13.88	47.88	154.51	330.83	793.29	2336.9
SSE opt irreg	0	13.29	42.86	132.16	271.01	607.5	1730.6
SSE équirép. reg	0	1956	2821	4245	5501	7024	9087
SSE équirép. irreg	0	505.8	860.8	1535	2042.6	2761.5	4677.1

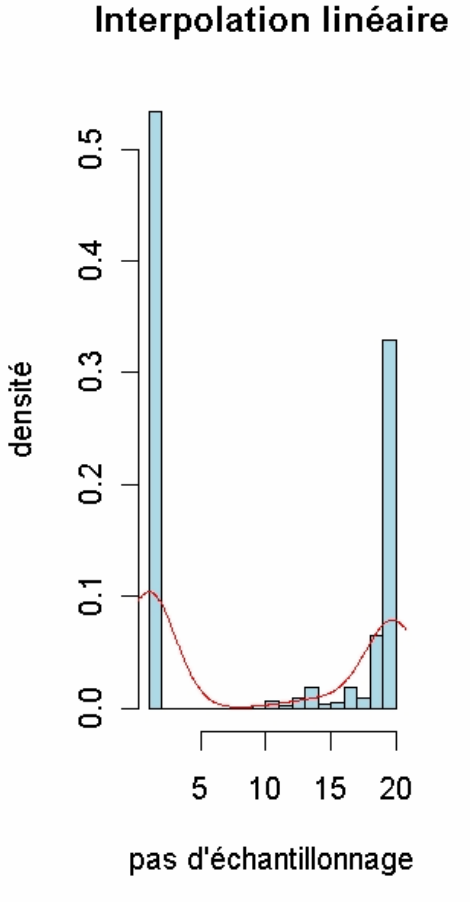
TAB. 1 – *tableau récapitulatif. SSE : Somme de l'écart quadratique. opt : pas d'échantillonnage (nb segments) obtenus par optimisation, et équirep. : même pas d'échantillonnage (nb segments) pour les CDCs. reg : échantillonnage régulier. irreg : échantillonnage par segmentation.*

Expérimentations (4/4)

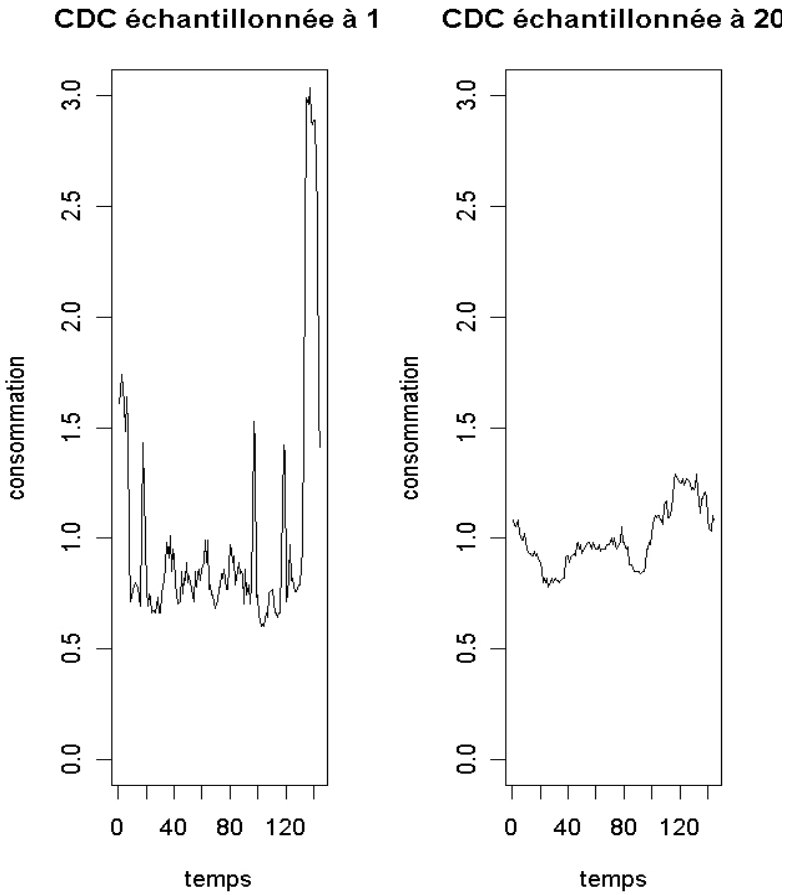


Distribution des pas d'échantillonnage

Expérimentations (4/4)



Distribution des pas d'échantillonnage



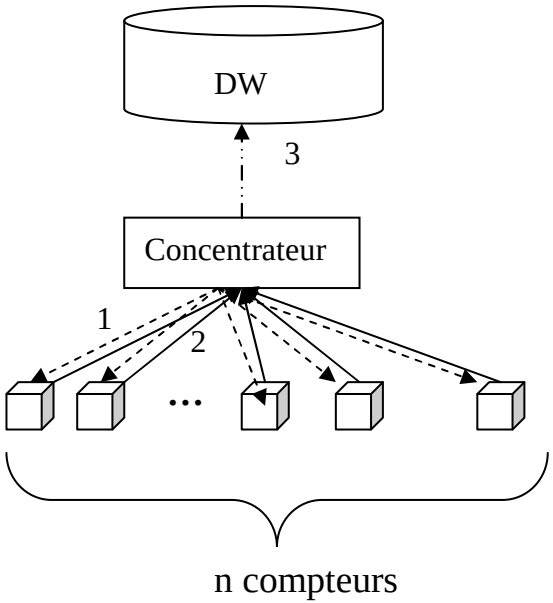
Deux exemples de CDCs échantillonnées différemment (1 et 20)

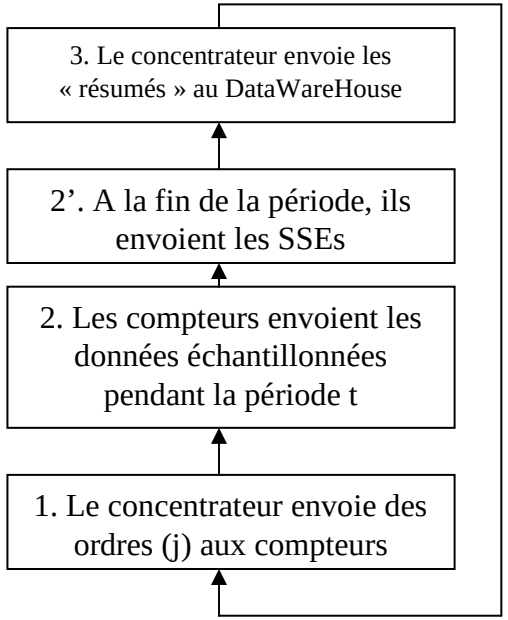
Plan de l'exposé

- Formulation du problème
- Méthode de résolution
- Expérimentations
- Traitement global du flux
- Travaux en cours

Traitement global du flux (1/4)

- Fenêtre t=0:
  - Compteurs envoient $\left\lfloor \frac{s}{n} \right\rfloor$ données
- Fenêtre t=1:
  - Les compteurs calculent SSEs et les envoient au concentrateur pour mettre à jour $W_{n*m'}$
  - Envoi aux compteurs les j_i optimisés
  - Le compteur applique le pas d'échantillonnage affecté par optimisation
- Fenêtre t=2: *idem que Fenêtre 1*
- …



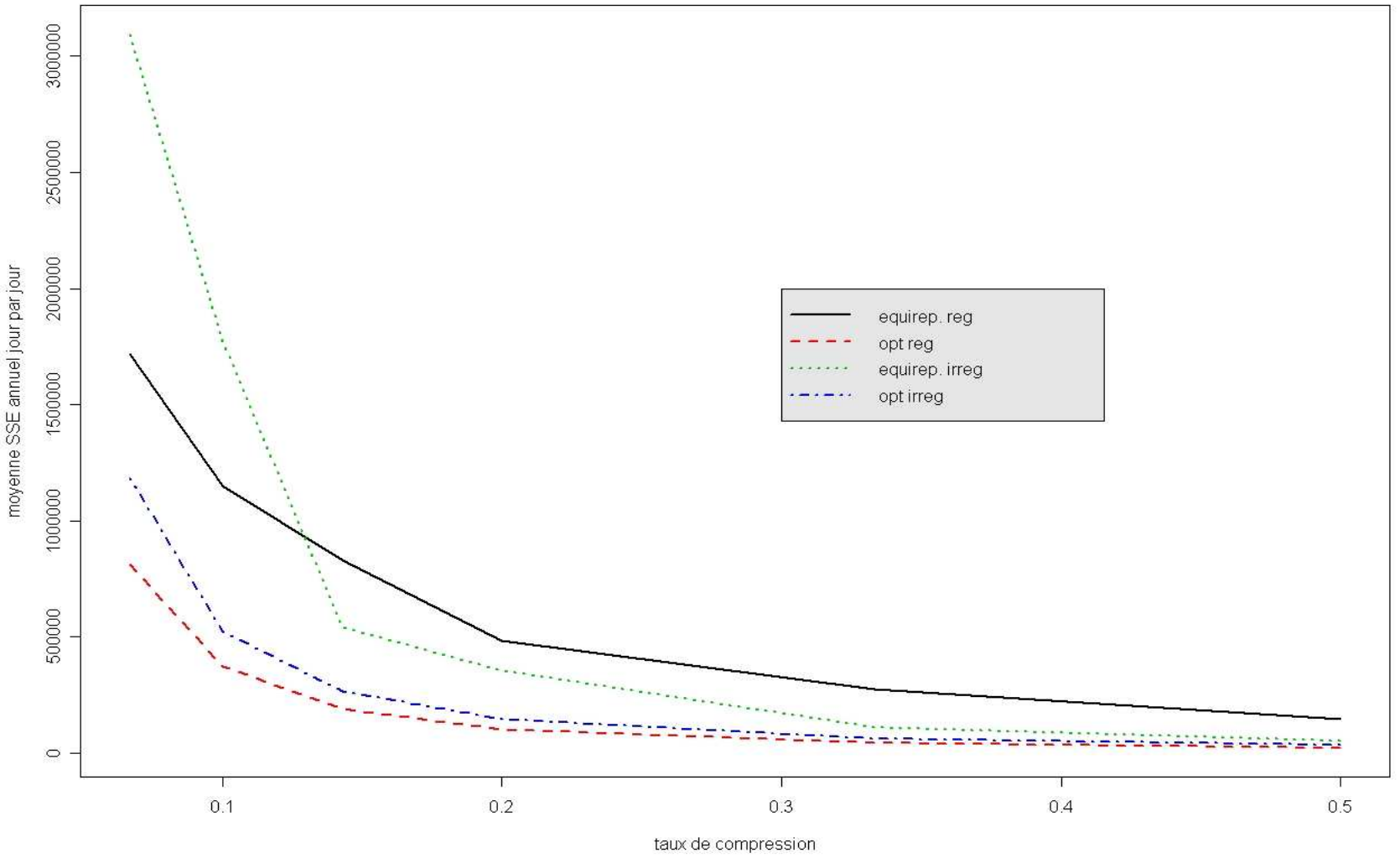


Traitement global du flux (2/4)

- 140 courbes relevées à un pas de temps de 30’ sur une année (365 jours), m=7

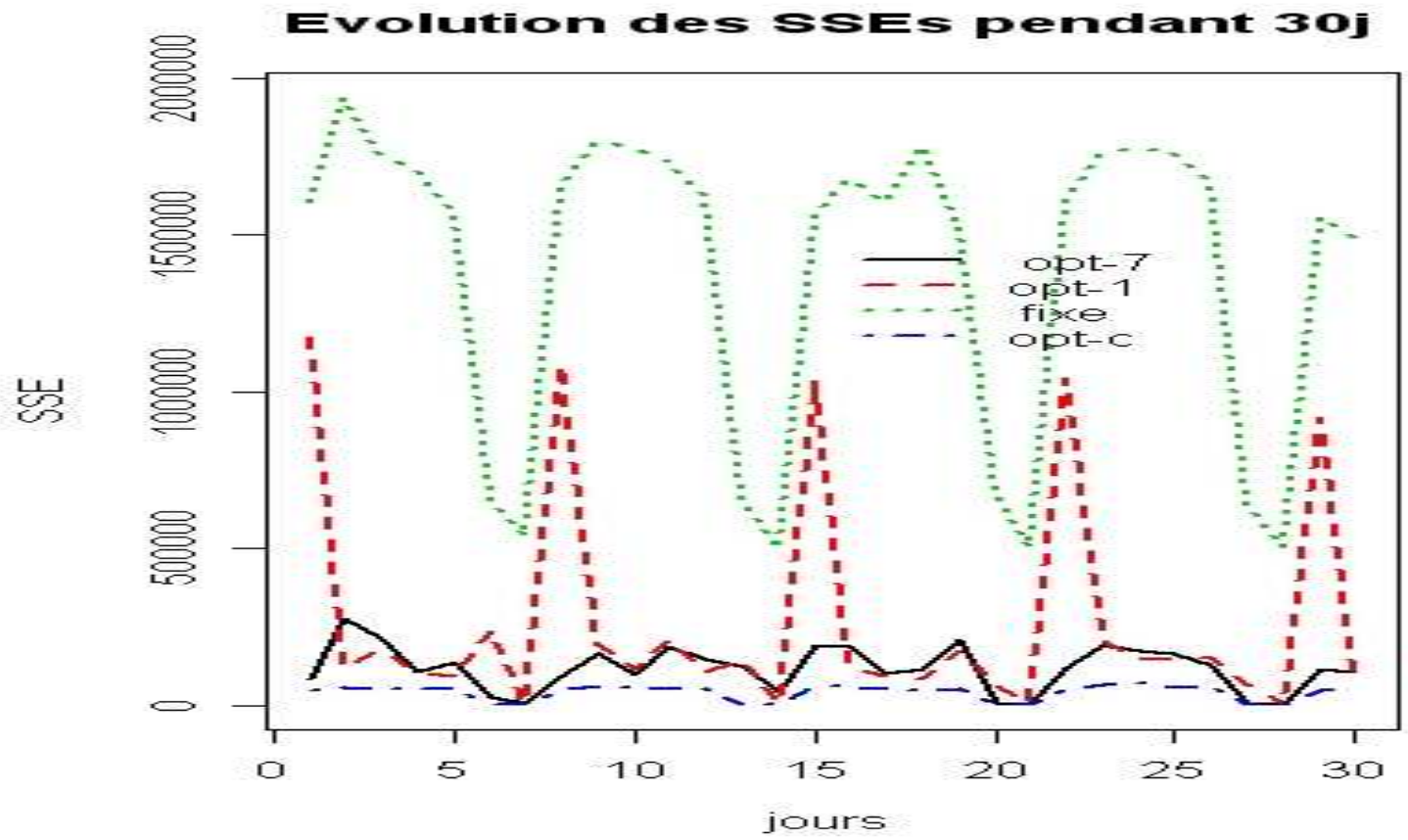
- Fenêtre temporelle = 1 journée
  - Phase d’optimisation utilise les données de la journée précédente et les pas d’échantillonnage sont appliqués à la journée courante

Traitement global du flux (2/4)



Traitement global du flux (3/4)

Fenêtre temporelle = 1 journée



Plan de l'exposé

- Formulation du problème
- Méthode de résolution
- Expérimentations
- Traitement global du flux
- Travaux en cours

Travaux en cours

- Utiliser une mesure d'erreur autre que la SSE: normes L_1, L_∞
- Agrandir le jeu de données: 1000 courbes correspondant à une période d'un an
- Suivre un panel de capteurs dans le temps Vs. Échantillonnage des mesures provenant de l'ensemble des capteurs pour répondre à des besoins d'analyse

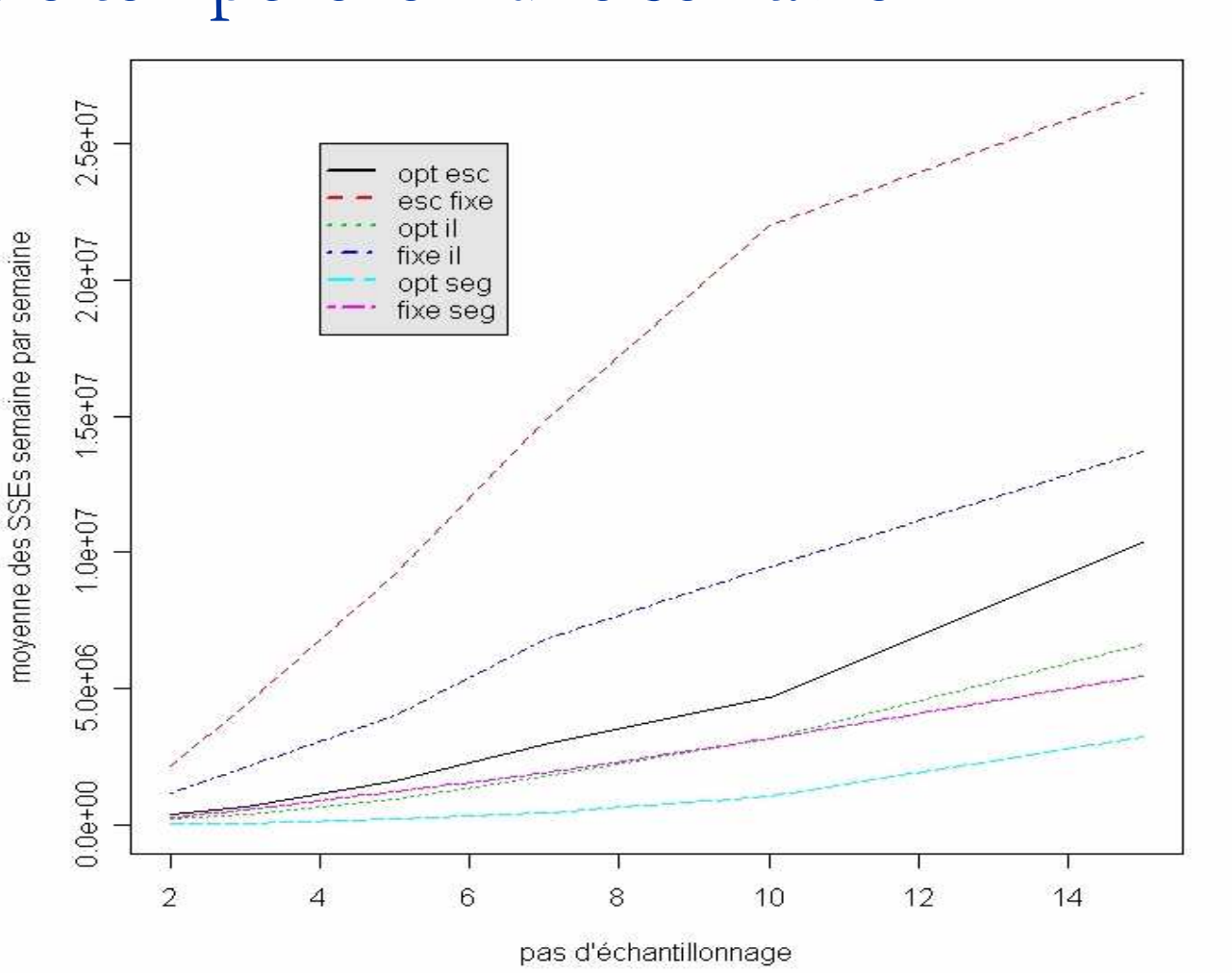
Merci pour votre attention

Raja CHIKY et Georges HEBRAIL

GET-ENST, EDF R&D

Traitement global du flux (4/4)

Fenêtre temporelle= une semaine



Traitement global du flux (4/4)

semaine vs. journée