

Data Science pour les données de capteurs

Georges HEBRAIL

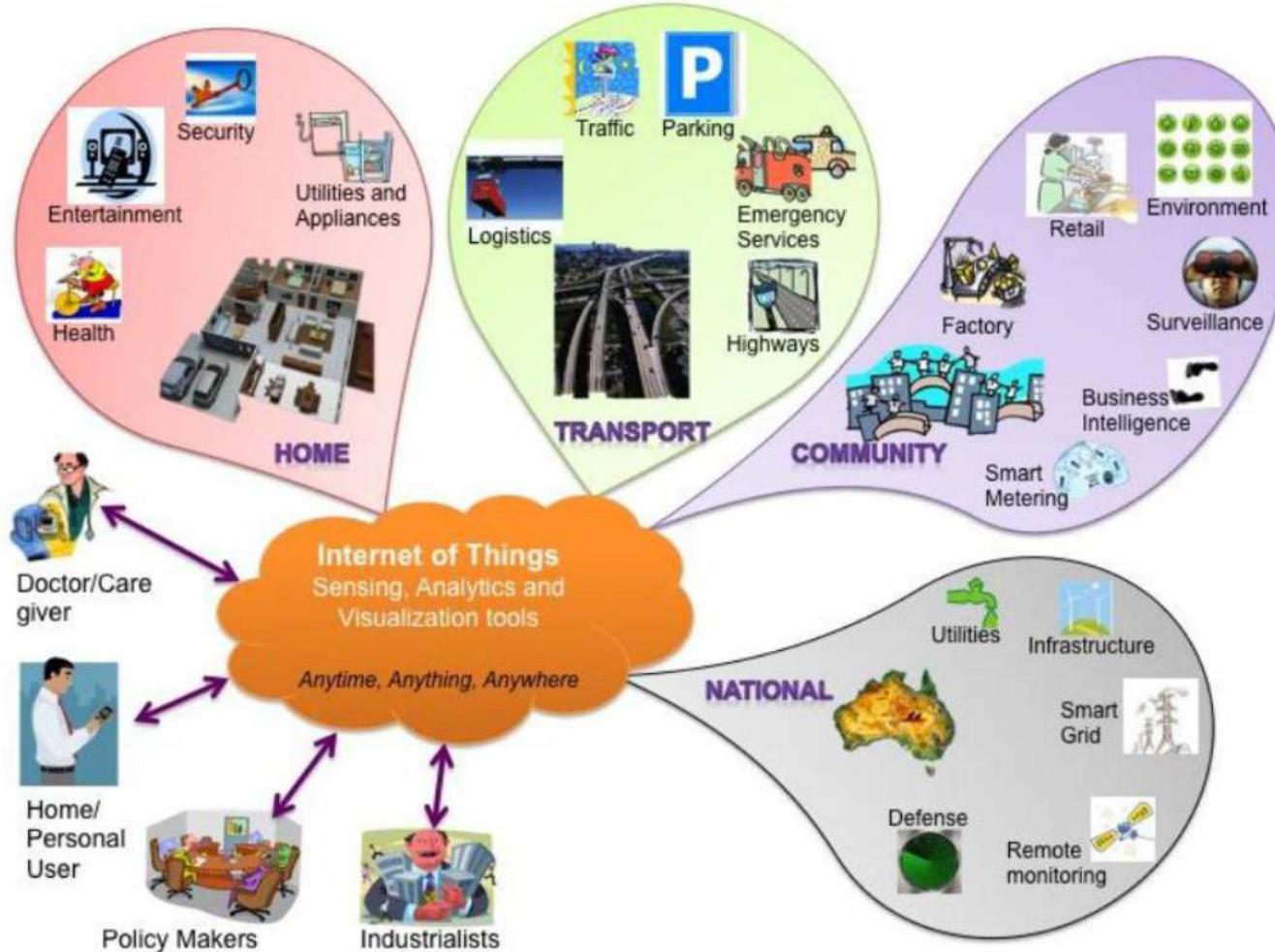
EDF R&D

9 Janvier 2017

AGENDA

- Exemples de motivation
- Données brutes de capteurs
- Préparation des données
- Similarités entre séries temporelles
- Analyse de séries temporelles
 - Analyse exploratoire
 - Analyse prédictive
- Etude de cas

Exemples de motivation



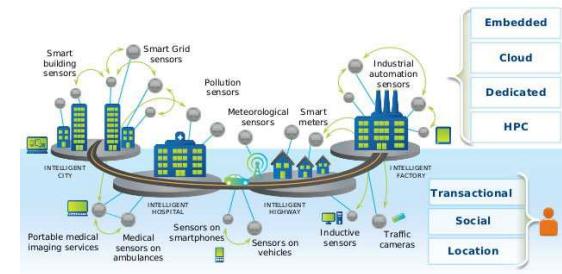
Exemples de motivation

- Capteurs dans l'environnement
(température, pluie, bruit, pollution, ...)



- Capteurs dans la ville
(transports, caméras, ...)

Smart City Sensor Model



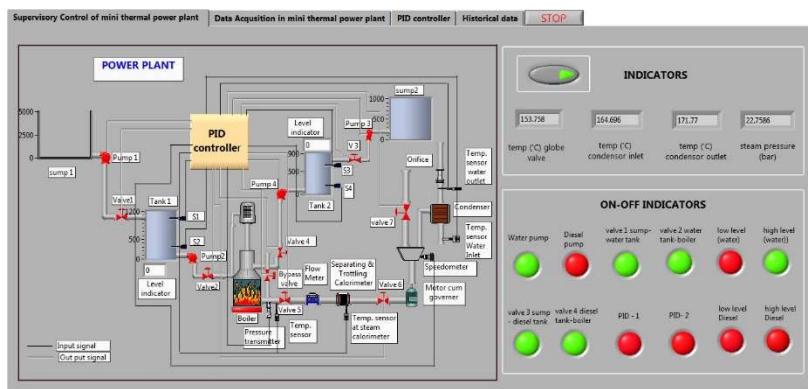
22

- Capteurs dans les bâtiments/maisons
(température, présence, compteur, ...)



Exemples de motivation

- Données de capteurs dans l'industrie
(production, machines, ...)



- Données de capteurs dans la santé
(mesures physiologiques, ...)



AGENDA

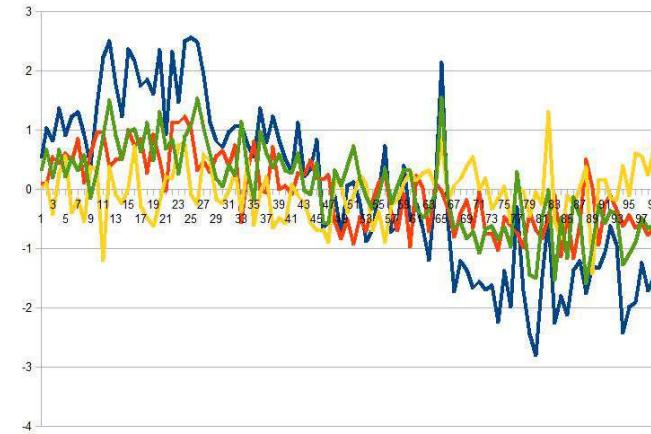
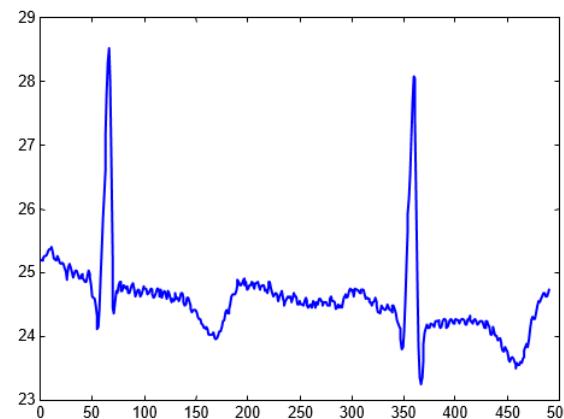
- Exemples de motivation
- Données brutes de capteurs
- Préparation des données
- Similarités entre séries temporelles
- Analyse de séries temporelles
 - Analyse exploratoire
 - Analyse prédictive
- Etude de cas

Données brutes de capteurs

- Données statiques :
 - Liées à l'équipement/composant/capteur : ID, modèle, année, constructeur, spécifications, ...
 - Liées à l'utilisateur/client : ID, type, contrat, localisation, ...
- Données dynamiques
 - Données de maintenance
 - Données continues
 - Données qualitatives

Données brutes de capteurs

- Données dynamiques continues (séries temporelles continues)
 - Capteurs mesurant des grandeurs continues physiques (environnement, électrique, mécaniques, ...)
 - MESURE (sensor ID, timestamp, [location], value)
 - Timestamp:
 - Instant : DD-MM-YYYY-HH-MM-SS (ou moins précis)
 - Mono/Multi-dimensionnel, Mono/Multi-varié



Données brutes de capteurs

- Données dynamiques de fonctionnement/maintenance/intervention
 - ID ticket, dates début/fin, équipement concerné, diagnostic, pièces changées, coût, technicien, ...

	Equipment	Notification Number	Notification date	Notification time	Text short	Ausfalldatum ab	Ausfallzeit ab	Ausfalldatum bis	Ausfallzeit bis	Division IPB	Material	IBCKF_M	IBCKF_MVKZ
23												5,263	5,071,584.67
24													3,301
25	14003	728100456470	07.09.2007	12:41:33	GENERATOR E	07.09.2007	12:41:33	07.09.2007	23:00:00	01	AXA	05764506	MEGALIX Cat.125
26		728100543184	21.02.2008	15:38:47	SOFTWARE A	21.02.2008	15:38:47	06.03.2008	22:38:34	01	AXA	04776063	HR Heliflex Optik o
27												05896852	CCR Board D71
28												07149979	Kamera Head TH 8
29		728100546276	27.02.2008	11:26:07	SOFTWARE H	27.02.2008	11:26:07	22.03.2008	02:56:31	01	AXA	07128866	CPU_D10_IS_OPE
30		728100963980	15.12.2009	11:22:05	X-RAY NO X-R	15.12.2009	11:22:05	15.12.2009	20:00:00	01	AXA	05997817	PCA QUAD DSP1
31	14004	400101791908	19.01.2007	16:38:12	CSE/27068 RE	19.01.2007	16:38:12	25.01.2007	20:24:00	01	AXA	01192124	ABSTANDSHALT
32		400101828175	28.02.2007	11:50:16	NO XRAY:RO	28.02.2007	11:50:16	29.06.2007	16:30:00	01	AXA	04775990	D90 Ein/Aus (Log)
33												05246264	CBL Signl.Comm D
34												05998377	ASM Host 3 Modu
35												07325579	ASM IMPAC Mast
36												07325900	ASM IPS2 WITH M
37												07326080	PCA Power ON 2
38		400101917799	05.06.2007	10:00:09	B-PLANE CAR	05.06.2007	10:00:10	09.06.2007	13:10:00	01	AXA	06465590	Potentiometer 354
39												10140940	Schaltleiste Bx/dB
40		400101950664	11.07.2007	17:56:41	CD WRITER DC	11.07.2007	17:56:41	26.07.2007	16:40:00	01	AXA	01768535	WANDLER;AC-DC
41												10051975	Adapter IDE to US
42												10051978	Drive DVD/CD R/W
43		400102027886	03.10.2007	12:24:11	A-PLANE FLU	03.10.2007	12:24:11	04.10.2007	09:40:00	01	AXA	07721603	FS Mono Artis Tis
44		400102247951	05.03.2008	17:06:54	ECC RAIL CLA	05.03.2008	17:06:54	06.03.2008	11:30:00	01	AXA	04775735	Zubehoer-Schiene
45		400102260600	13.03.2008	10:33:50	TUBE STARTU	13.03.2008	10:33:51	14.03.2008	10:40:00	01	AXA	07124139	Anlassgeraet N75
46		400102598104	21.10.2008	15:57:59	CSE COLE REF	21.10.2008	15:57:59	28.11.2008	09:10:00	01	AXA	04776063	HR Heliflex Optik o

Source : "Predictive Maintenance in a Machine Learning Perspective", IEEE BigData 2015 Tutorial", Zhuang (John) Wang.

Données brutes de capteurs

- Données dynamiques qualitatives
(séries temporelles qualitatives)
 - Enregistrement d'événements (ex. logs)
 - EVENT (ID équipement, timestamp, code événement, message)

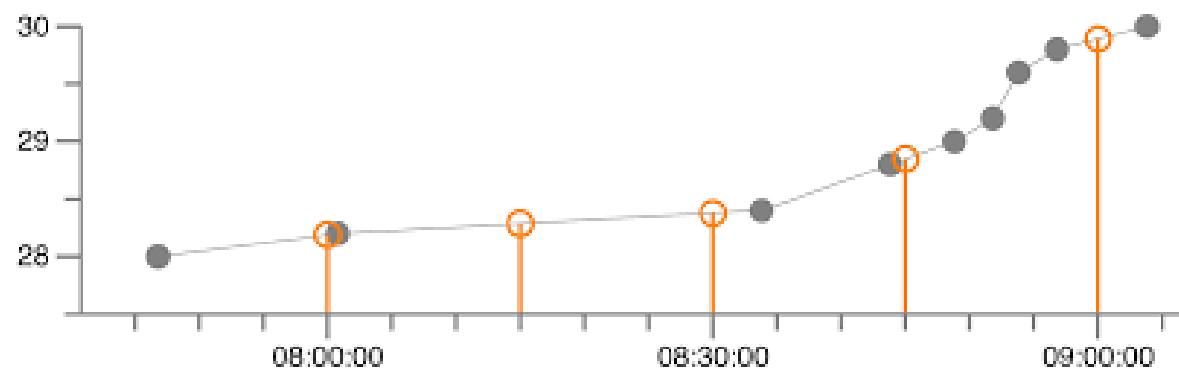
Time stamp	Event code	Message
2012-09-08 21:21:09	CT_IIS	IRS scan information: Topo tot Rdg: 6460, defects: HV-Drop 0, HT
2012-09-08 21:21:07	CT_IS	320 First data from IRS arrived in IS
2012-09-08 21:21:06	CT_CHR	107 SKIP button is hidden (bolus exam not started active entry is not n
2012-09-08 21:21:06	CT_CHR	107 CANCEL button is dimmed (scanstate == UiScanning)
2012-09-08 21:21:06	CT_CHR	107 CANCELMOVE button is hidden (scanstate == UiScanning)
2012-09-08 21:21:06	CT_STC	93 Zero button has been locked.
2012-09-08 21:21:06	CT_ISV	49 Timer was started waiting for some answer from IRS Recc
2012-09-08 21:21:06	CT_CHR	107 Button(s) disabled:#CLOSE PATIENT#EXAM#PATIENT REGISTRATION EXAM#
2012-09-08 21:21:06	CT_ISV	49 Timer was started waiting for SCAN_DONE message for 77E
2012-09-08 21:21:06	ACU	3037 Control info ACU (E c0 03 25 d4 4a 00 00)
2012-09-08 21:21:06	ACU	3037 Control info ACU (E c0 03 25 d3 7e 00 00)
2012-09-08 21:21:06	TG	9 A new planning image of type CLEAR SEGMENT with series loid EMPTY
2012-09-08 21:21:06	MSM	211 (+) Receiving MeasStart request (scan 0 of range 0)
2012-09-08 21:21:06	SSQ	493 received IS-Notification: SsqMeasISClient::onIrsStarted / IS_MsgId
2012-09-08 21:21:02	CT_SIS	120 eStart button released&
2012-09-08 21:21:02	CT_SIS	85 Multinlexer: the START-button on control box has been released.
2012-09-08 21:21:02	CT	Semi-structured text (high- cardinality variable)
2012-09-08 21:21:02	CT	Cardinality variable
2012-09-08 21:21:02	CT_CHR	107 LOAD button is suspended
2012-09-08 21:21:02	CT_CHR	107 SUSPEND button is suspended
2012-09-08 21:21:02	CT_CHR	107 Button(s) disabled:#CLOSE PATIENT#EXAM#PATIENT REGISTRATION EXAM#
2012-09-08 21:21:02	CT_ISV	45 Sending IsvMsgStart message of size 72 Bytes to IRS.

AGENDA

- Exemples de motivation
- Données brutes de capteurs
- Préparation des données
- Similarités entre séries temporelles
- Analyse de séries temporelles
 - Analyse exploratoire
 - Analyse prédictive
- Etude de cas

Préparation des données

- Mise à des pas de temps réguliers
 - La plupart des méthodes d'analyse l'imposent
 - Mesures irrégulières
 - Mesures de plusieurs séries à des instants différents
 - L'interpolation linéaire est souvent une mauvaise solution
 - Respect des dérivées, de l'intégrale, ...



Préparation des données

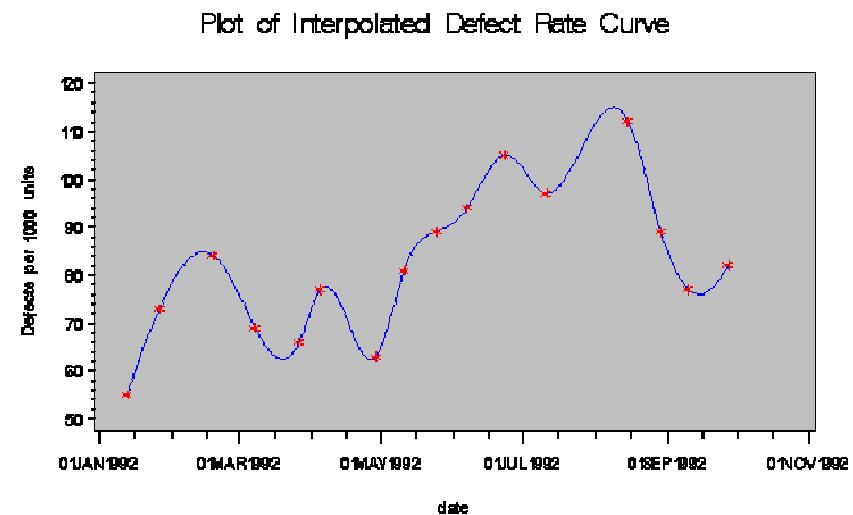
- On « plonge » les données dans un espace fonctionnel (lissage)
 - Détermination de l'espace fonctionnel
 - Ré-échantillonnage au pas souhaité

$$y_j = x(t_j) + \epsilon_j$$

(y observé, x fonction, ϵ erreur)

$$x(t) = \sum_{k=1}^K c_k \Phi_k(t)$$

- Choix des fonctions de base Φ_k
 - Bases de Fourier
 - Bases polynomiales
 - Bases de splines (régularisées)



Préparation des données

- Utilisations d'un espace fonctionnel
 - Utilisation des valeurs ré-échantillonnées (cf. pas de temps réguliers)
 - Si plusieurs individus dans le même espace : utilisation des coordonnées dans l'espace
- Détection/correction des valeurs manquantes/aberrantes
 - Utilisation d'une représentation fonctionnelle
 - Valeur manquante sur courte durée : interpolation
 - Valeur manquante sur une plus longue durée (ex. 1j) : recopie de la veille

AGENDA

- Exemples de motivation
- Données brutes de capteurs
- Préparation des données
- **Similarités entre séries temporelles**
- Analyse de séries temporelles
 - Analyse exploratoire
 - Analyse prédictive
- Etude de cas

Similarités entre séries temporelles

- Utile dans de nombreuses situations

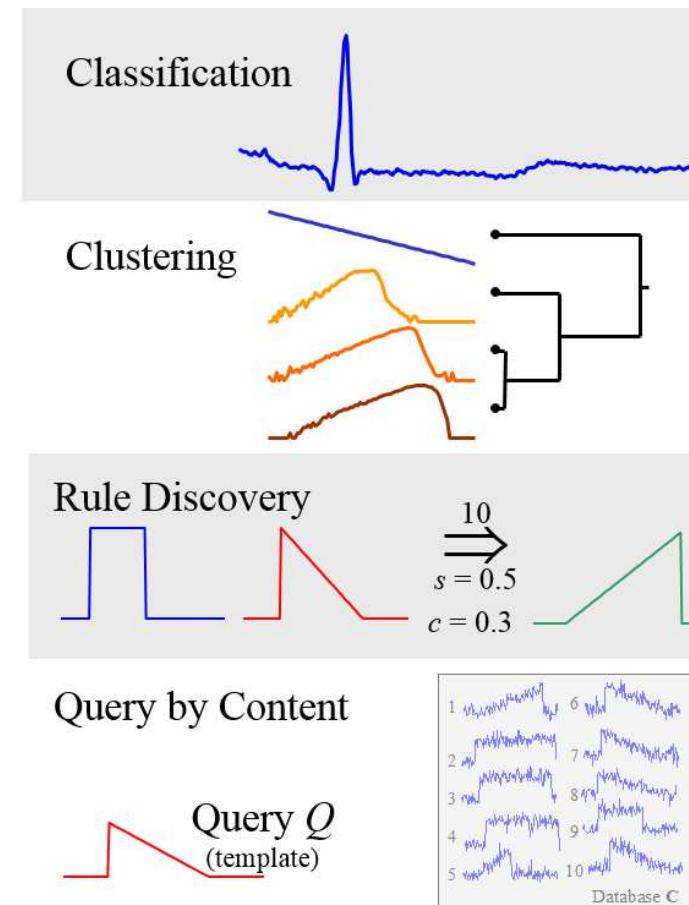
Time Series Similarity

Defining the similarity between two time series is at the heart of most time series data mining applications/tasks

Thus time series similarity will be the primary focus of this tutorial.

Source : "Mining time series", CS240B Notes by Carlo Zaniolo

Source: "Mining time series", CS240B Notes by Carlo Zaniolo, UCLA CS Dept



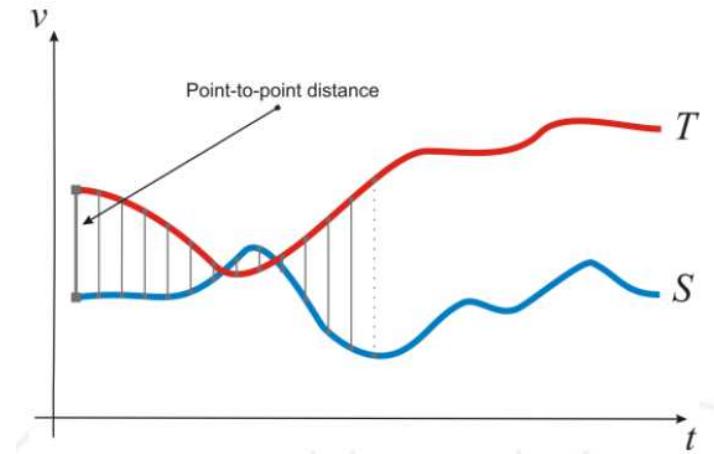
Similarités entre séries temporelles

- Métriques et similarités
- Mesures point à point
 - Minkowski (euclidienne – 2, Manhattan – 1, Maximum - ∞)

$$d(T, S) = \sqrt[p]{\sum_{i=1}^n (T_i - S_i)^p}$$

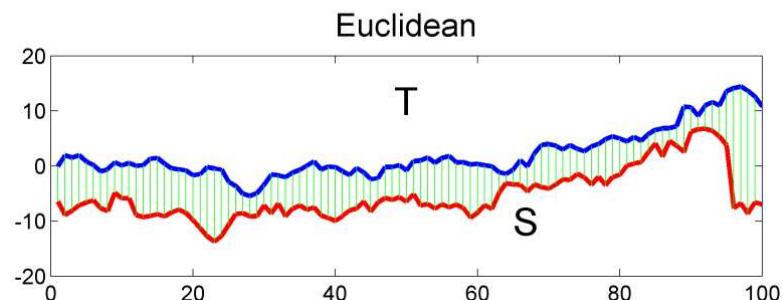
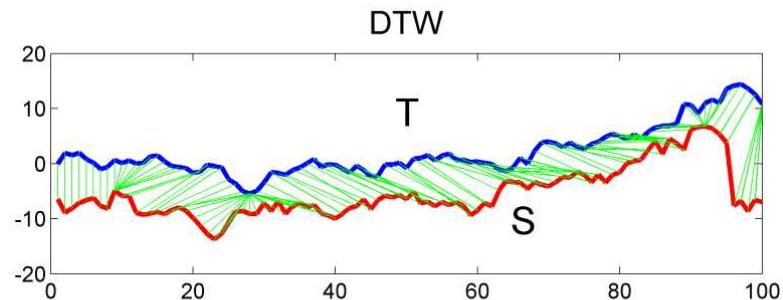
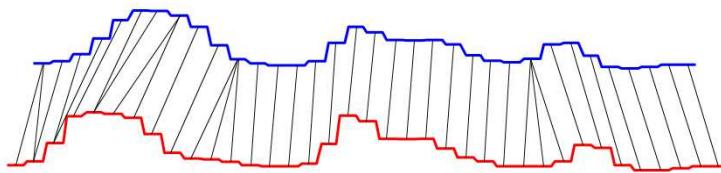
- Corrélation

$$r_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_{i-l} - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (y_{i-l} - \bar{Y})^2}}$$



Similarités entre séries temporelles

- Dynamic Time Warping (DTW)
 - Distorsion temporelle
 - Programmation dynamique



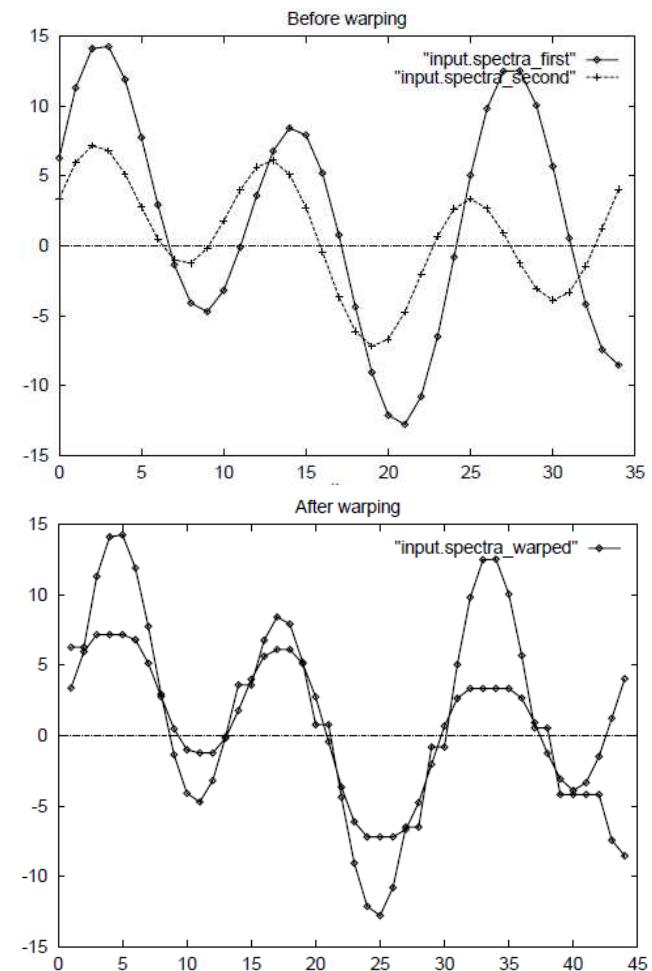
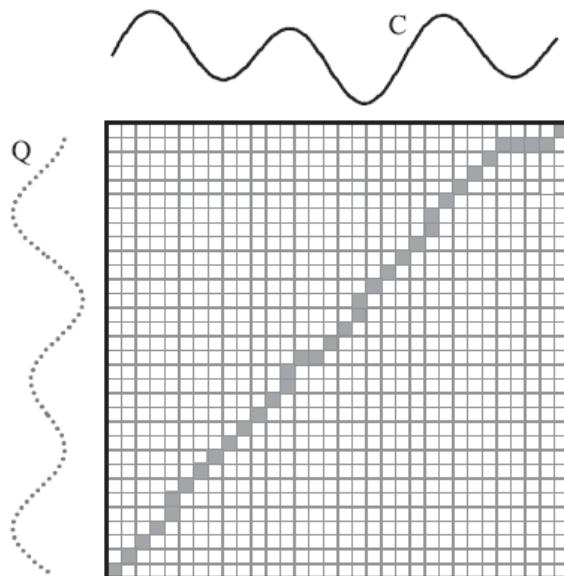
$$DTW(T, S) = \min \left(\sqrt{\sum_{k=1}^K w_k} \right)$$

The warping path is subject to several constraints . Given $w_k = (i, j)$ and $w_{k-1} = (i', j')$ with $i, i' \leq n$ and $j, j' \leq m$:

1. **Boundary conditions.** $w_1 = (1,1)$ and $w_K = (n, m)$.
2. **Continuity.** $i - i' \leq 1$ and $j - j' \leq 1$.
3. **Monotonicity.** $i - i' \geq 0$ and $j - j' \geq 0$.

Similarités entre séries temporelles

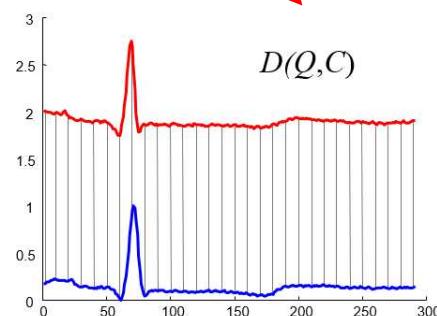
- Dynamic Time Warping (DTW)
 - Distorsion temporelle
 - Programmation dynamique



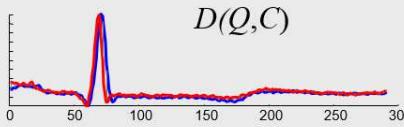
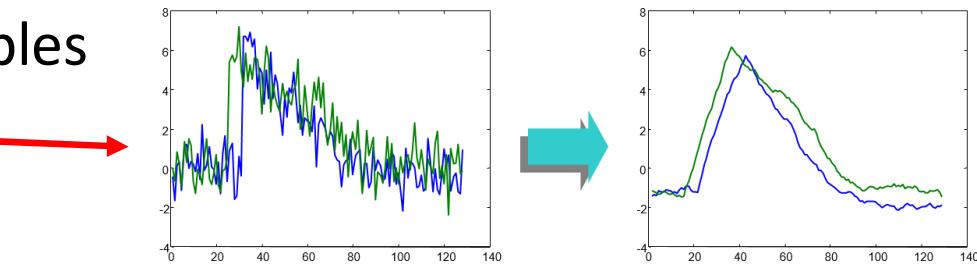
Similarités entre séries temporelles

- Transformations préalables

- Débruitage
- Amplitude
- Niveau



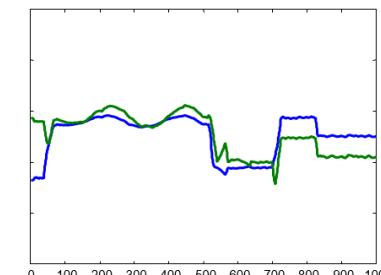
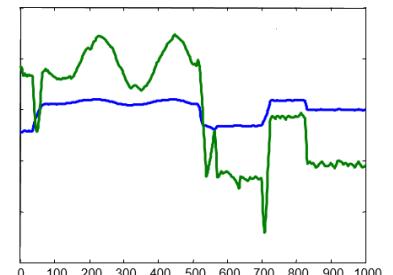
$Q = Q - \text{mean}(Q)$
 $C = C - \text{mean}(C)$
 $D(Q, C)$

$$Q = \text{smooth}(Q)$$

$$C = \text{smooth}(C)$$

$$D(Q, C)$$



$Q = (Q - \text{mean}(Q)) / \text{std}(Q)$
 $C = (C - \text{mean}(C)) / \text{std}(C)$
 $D(Q, C)$

AGENDA

- Exemples de motivation
- Données brutes de capteurs
- Préparation des données
- Similarités entre séries temporelles
- Analyse de séries temporelles
 - Analyse exploratoire
 - Analyse prédictive
- Etude de cas

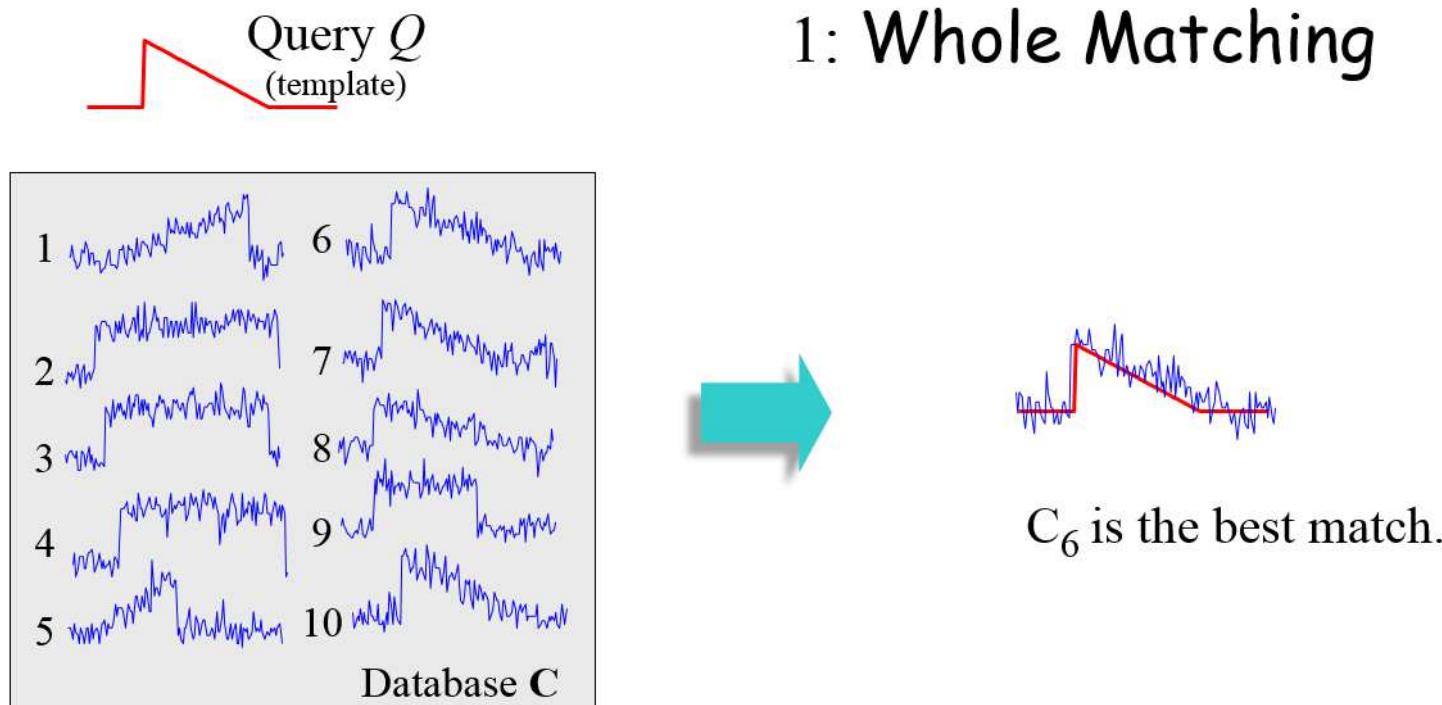
Analyse des séries temporelles

- « Data analytics »
 - Exploratory analysis / unsupervised learning
 - Explorer, synthétiser des données
 - Predictive analysis / supervised learning
 - Utiliser des données pour prédire/prévoir

Analyse exploratoire de séries temporelles

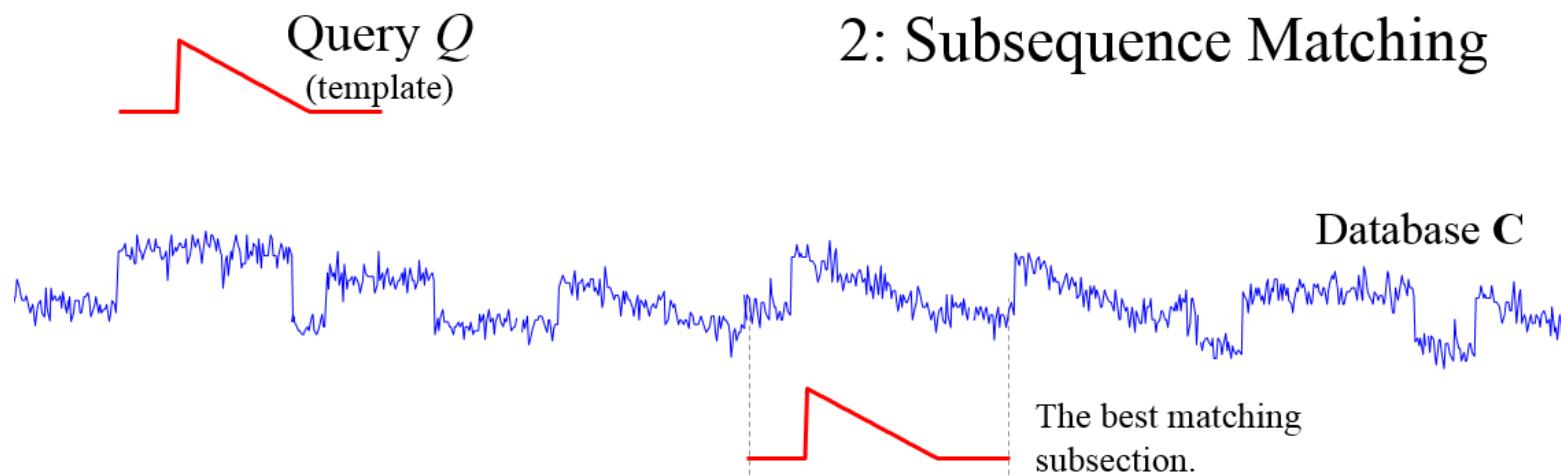
- Requêtage de séries temporelles
(Similarity based time series retrieval)
- Visualisation
- Classification automatique (clustering)
- Recherche de motifs fréquents

Requête de séries temporelles



Source : "Mining time series",
CS240B Notes by Carlo
Zaniolo

Requête de séries temporelles

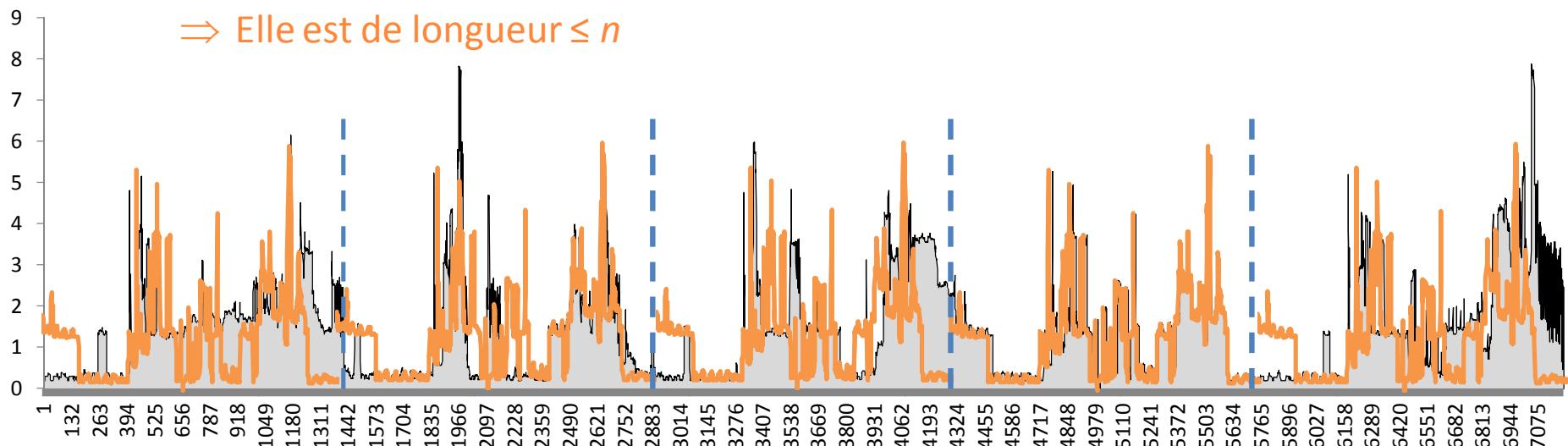


Given a Query Q , a reference database \mathbf{C} and a distance measure, find the location that best matches Q .

Requête de séries temporelles

- **Requête "sautante" (*jumping window*)**

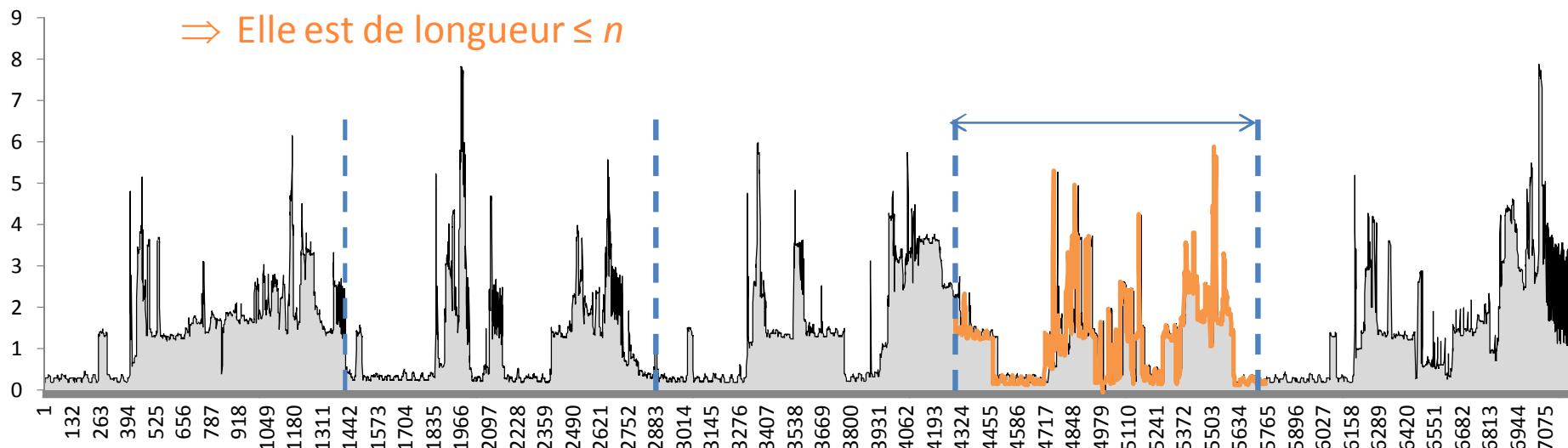
- ⇒ La requête "saute" de segment en segment
- ⇒ Les instants de la série-requête correspondent point à point à ceux de la série actuellement examinée
- ⇒ Elle est de longueur $\leq n$



Requête de séries temporelles

- **Requête "sautante" (*jumping window*)**

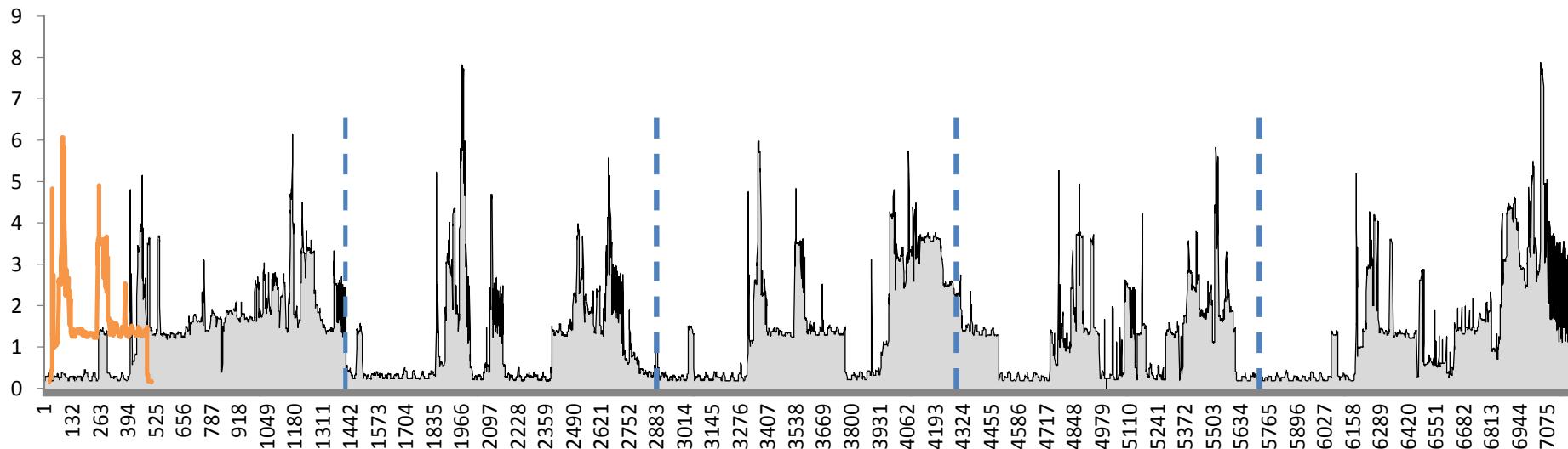
- ⇒ La requête "saute" de segment en segment
- ⇒ Les instants de la série-requête correspondent point à point à ceux de la série actuellement examinée
- ⇒ Elle est de longueur $\leq n$



Requête de séries temporelles

- **Requête "glissante" (*sliding window*)**

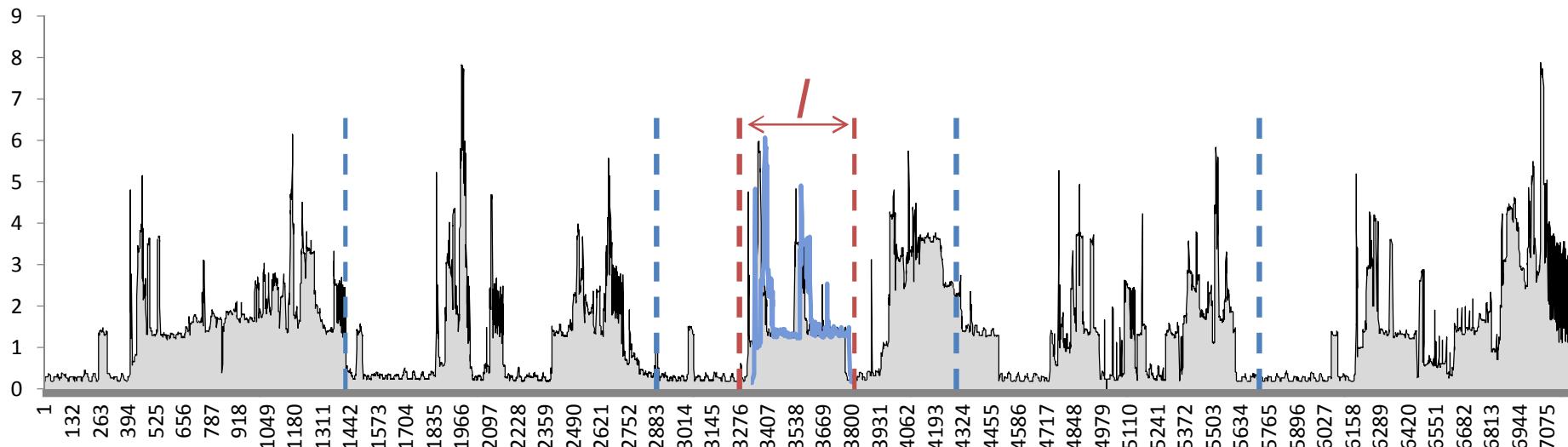
- ⇒ La requête "glisse" en avançant **point à point** sur chaque segment
- ⇒ Elle peut tomber à cheval sur 2 segments
- ⇒ Elle est de longueur l



Requête de séries temporelles

- **Requête "glissante" (*sliding window*)**

- ⇒ La requête "glisse" en avançant **point à point** sur chaque segment
- ⇒ Elle peut tomber à cheval sur 2 segments
- ⇒ Elle est de longueur l



Requête de séries temporelles

- Requête ***Top-K***

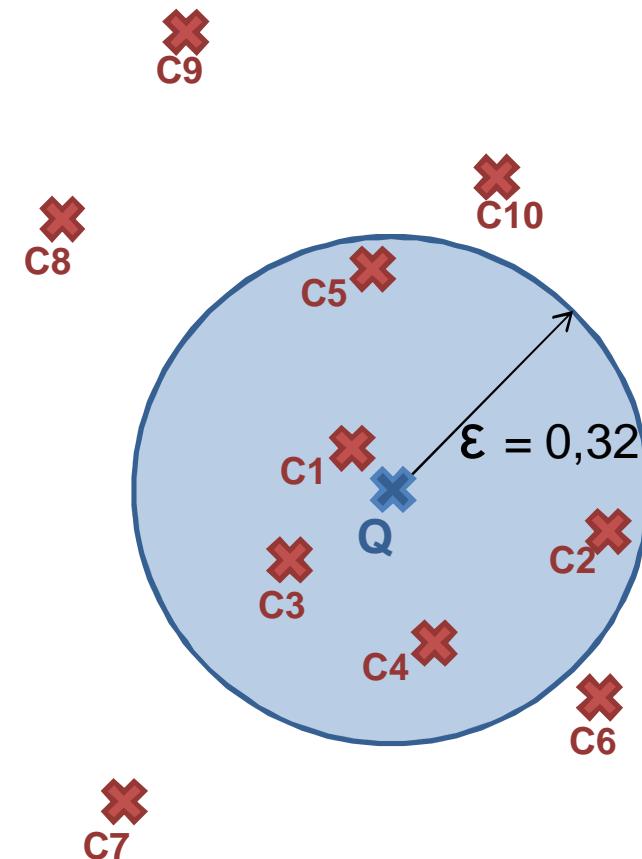
(K plus proches voisins)

Si $K = 3$
C1, C3, C4

- Requête ***Range-query (ϵ)***

(dans un rayon ϵ)

Si $\epsilon = 0,32$
C1, C2, C3, C4, C5

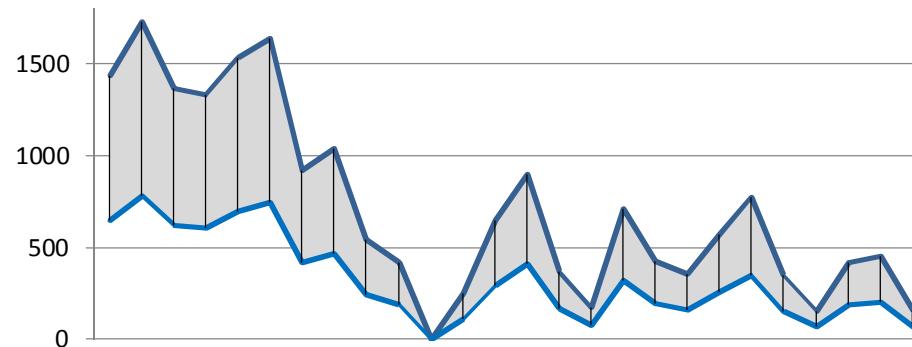


Requête de séries temporelles

- Difficultés
 - Sémantique de la recherche
 - Choix de la représentation, de la distance
 - Performance en temps de réponse
 - Parcours séquentiel parallélisé
 - Réduction de la dimension
 - Pour que des techniques d'indexation soient efficaces (R-trees)
 - Pour réaliser une pré-selection sur une représentation réduite

Requête de séries temporelles

- Représentation brute
- Représentation dans un espace plus approprié
 - **Normaliser**
⇒ Solution typique : diviser par la moyenne ou centrer-réduire

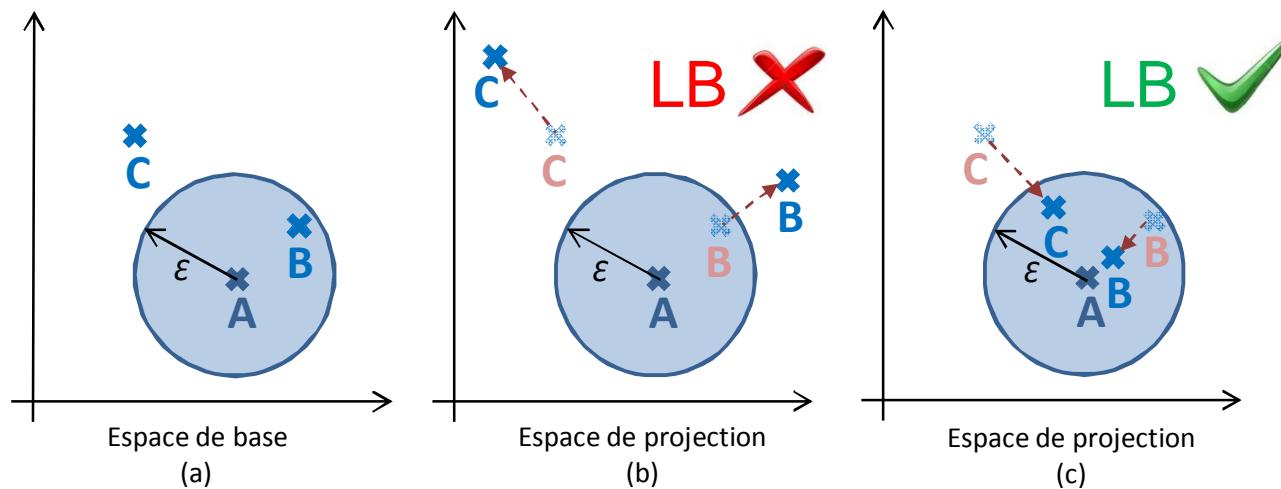


- Effectuer une transformation réduisant la dimension dans le but de :
- OU
 - ✓ De *lisser* la courbe et donc de modifier le critère de recherche
 - ✓ D'*optimiser* le temps d'exécution de la requête par une **présélection**
⇒ Fourier, ondelettes, SVD, ...

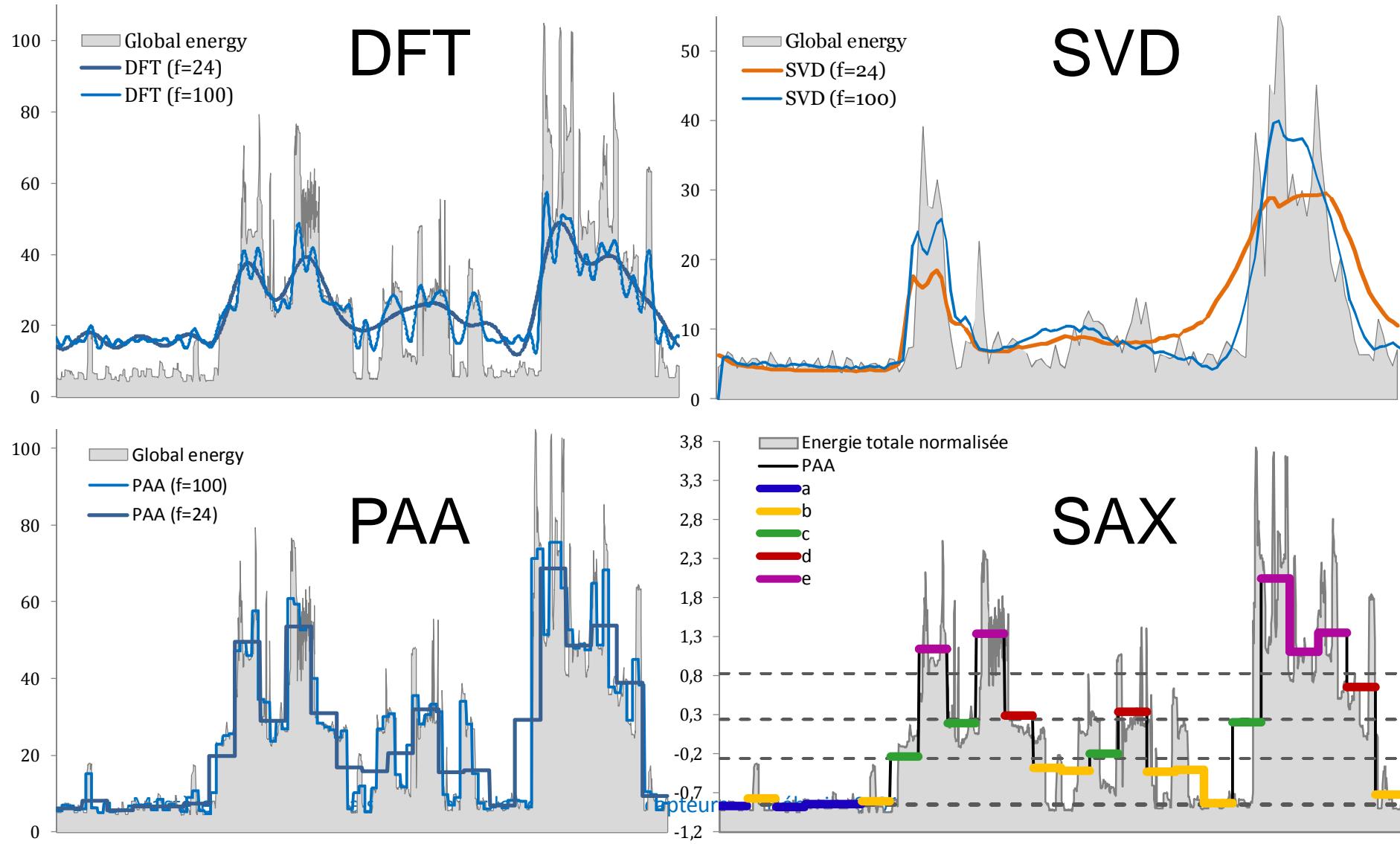
Requête de séries temporelles

- Comment réduire la dimension efficacement ?
 - Extraire de la série f caractéristiques ($f < n$) grâce à une fonction $F...$
 - ... vérifiant la propriété du "Lower Bounding" :

$$D_{\text{espace de projection}}(F(A), F(B)) \leq D_{\text{espace de base}}(A, B)$$

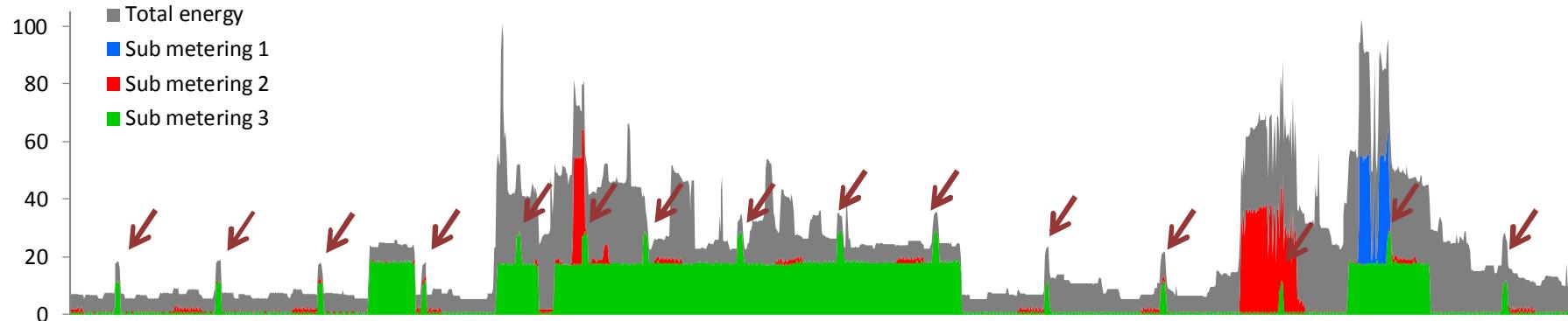


Requête de séries temporelles

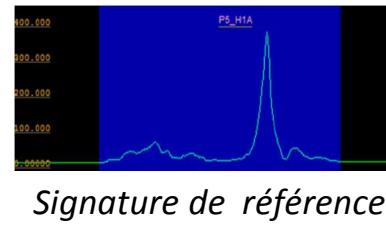


Exemples EDF

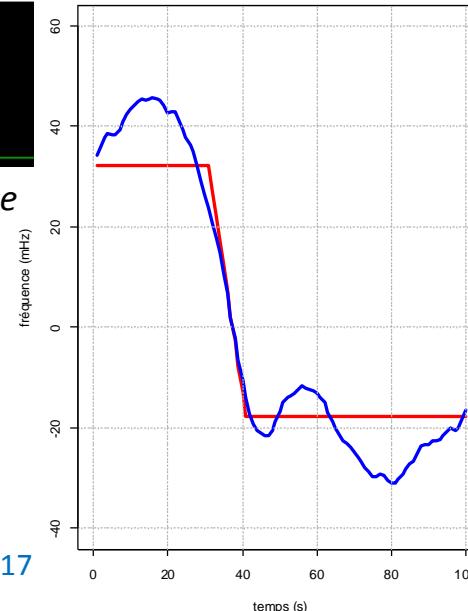
- Courbes de **consommation d'électricité**



- Aide au diagnostic :** rapprochement de signatures lors de transitoires

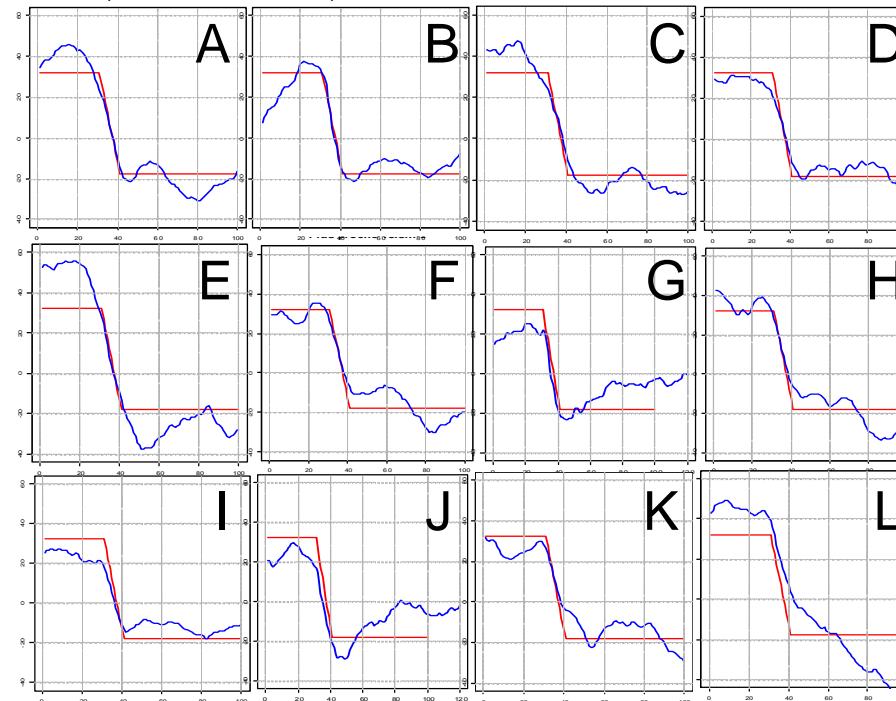


- Maintien de la fréquence** dans les centrales



Exemples EDF

- **Maintien de la fréquence par les centrales**

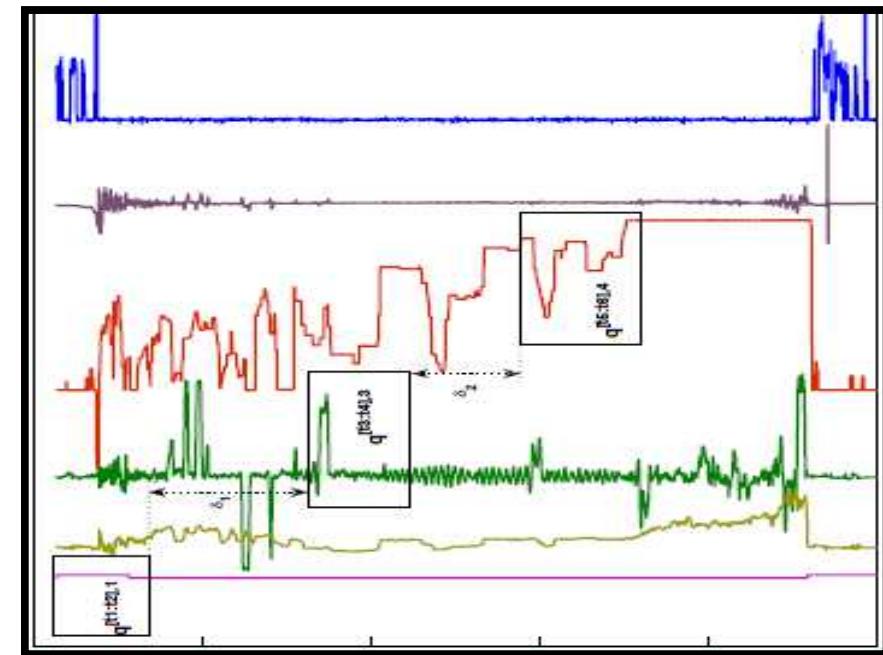
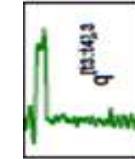
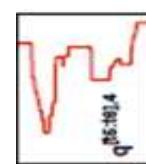


Requête de séries temporelles

- Séries multi-dimensionnelles

Rapprochement de signatures lors de transitoires

$Q =$

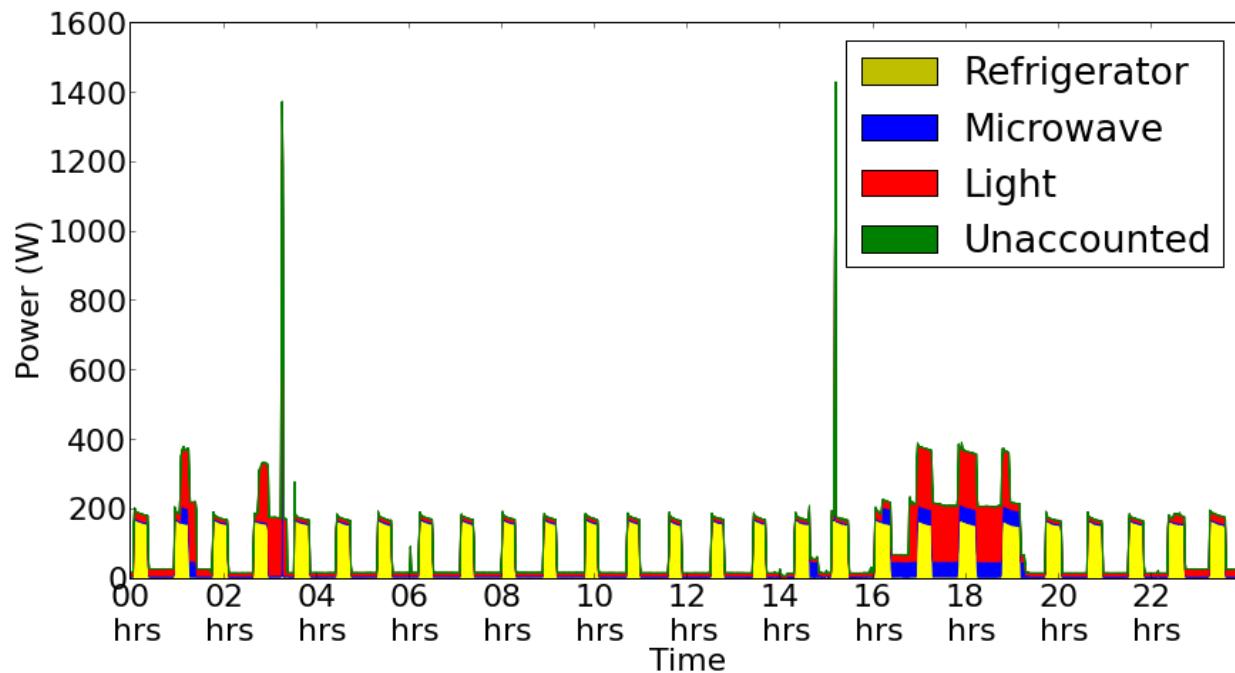


Analyse exploratoire de séries temporelles

- Requêtage de séries temporelles
(Similarity based time series retrieval)
- **Visualisation**
- Classification automatique (clustering)
- Recherche de motifs fréquents

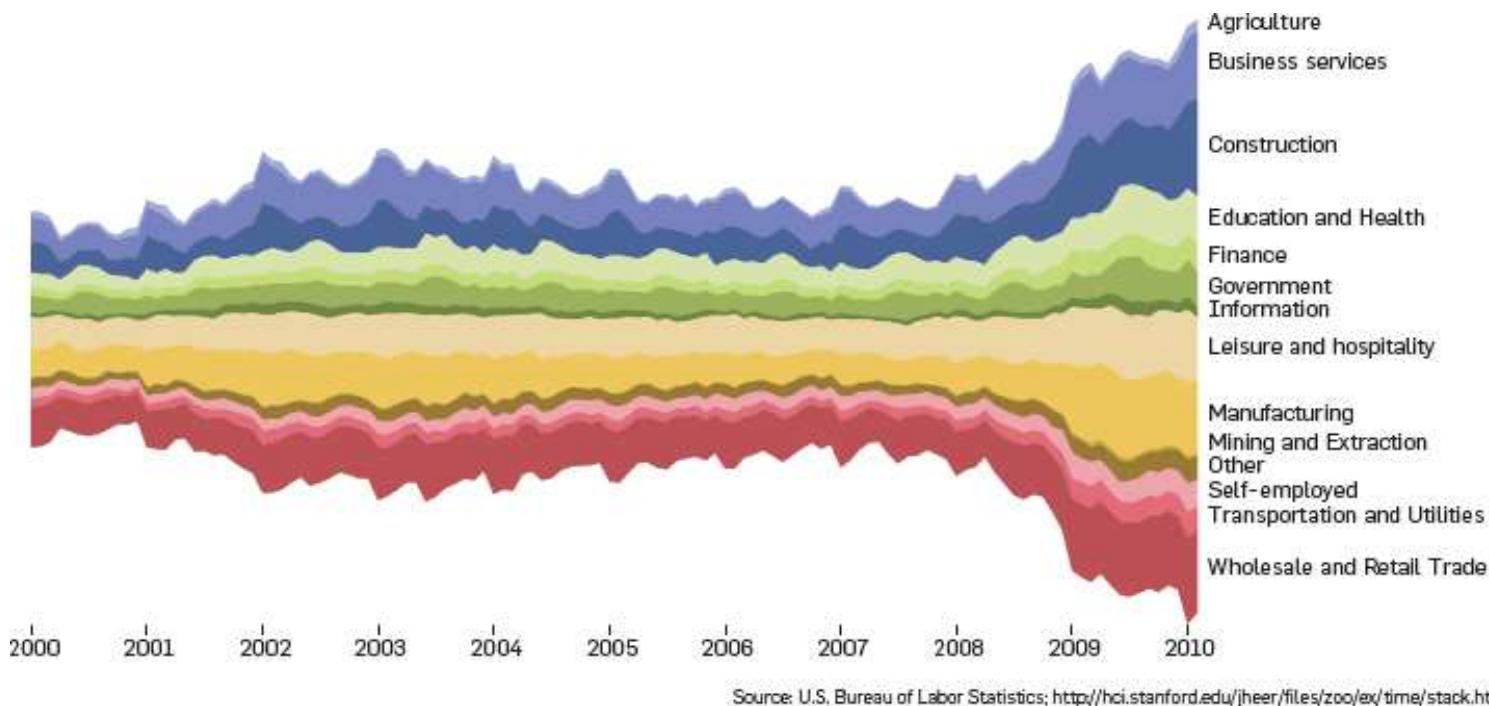
Visualisation de séries temporelles

- Naturellement la visualisation est adaptée
 - Mono-dimensionnelle, multi-dimensionnelle



Visualisation de séries temporelles

- http://cacm.acm.org/magazines/2010/6/92482-a-tour-through-the-visualization-zoo/fulltext?hc_location=ufi%3Fmobile%3Dtrue

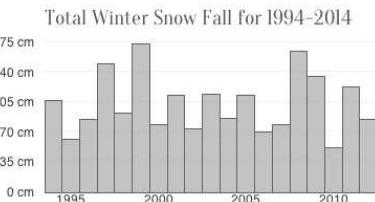
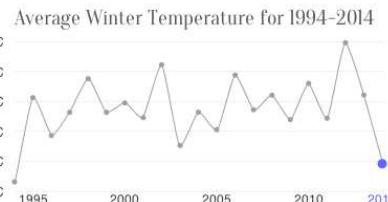


Visualisation de séries temporelles

<http://www.neoformix.com>

2014 was the Coldest Winter in Twenty Years for Markham

The average temperature was -8.2°C for the winter of 2014. That's the coldest since 1994 when the average temperature was -9.4°C during the winter.
Between Dec 21, 2013 and Mar 21, 2014 a total of 146 cm of snow fell in Markham. That makes it the snowiest winter since 2008 and the 4th snowiest in 20 years.

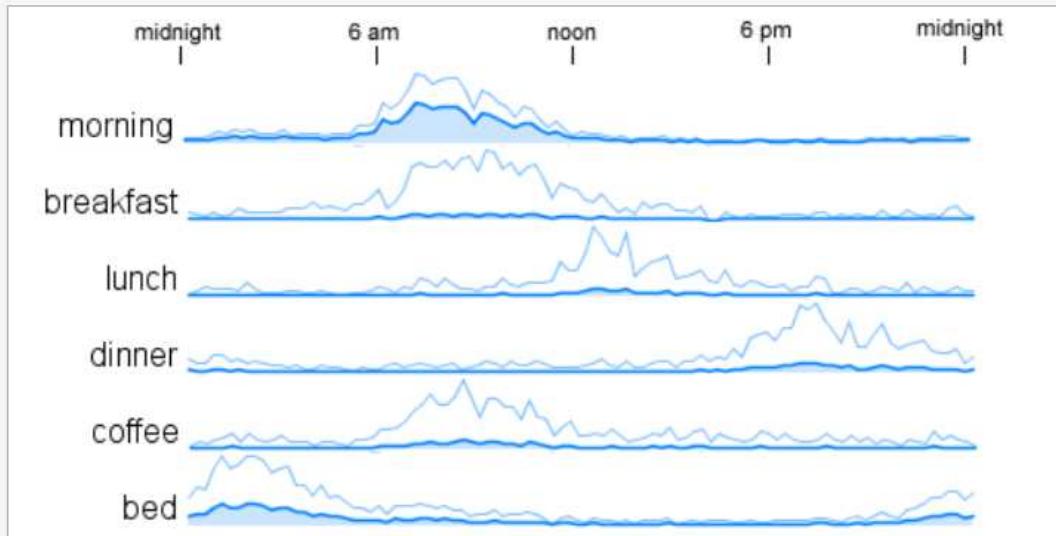


Analysis and Graphic by Jeff Clark - neoformix.com - © 2014, Data from Environment Canada for Buttonville Airport

Visualisation de séries temporelles

<http://www.neoformix.com>

In my last post, [Time Series for Word Counts in Tweets](#), I showed some graphs illustrating how often a word was used in tweets during the various times of day. I'm using the same data here, 575,962 tweets sent from the Toronto area in the month of July 2009. Some of the graphs show very similar shapes, for example 'morning', 'breakfast', and 'coffee' in the set below.



Visualisation de séries temporelles

- Graphes d'intensité
(heatmaps)

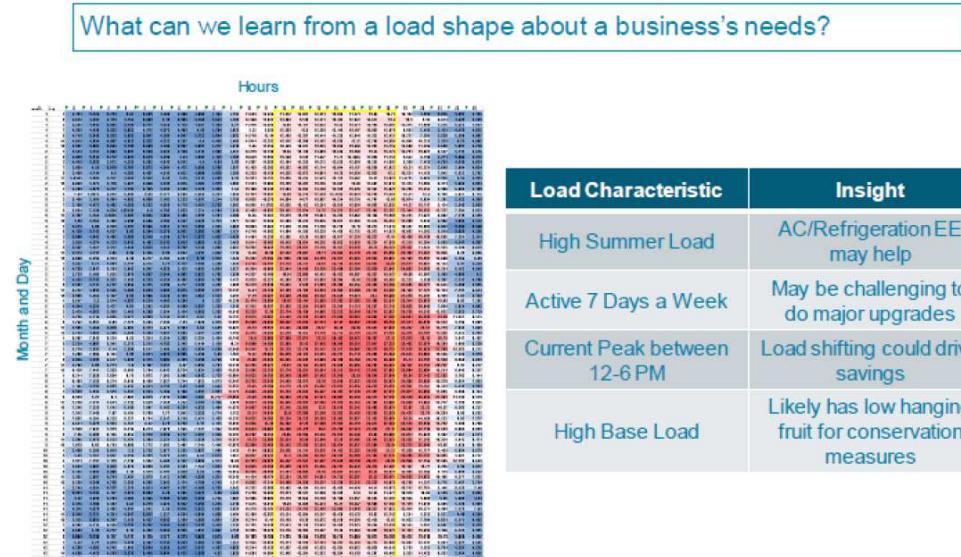
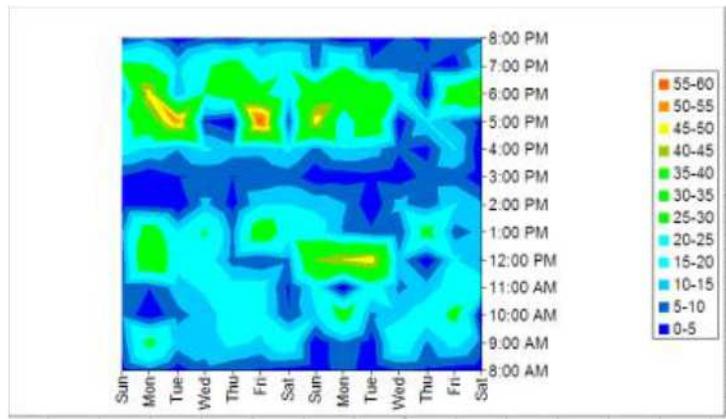
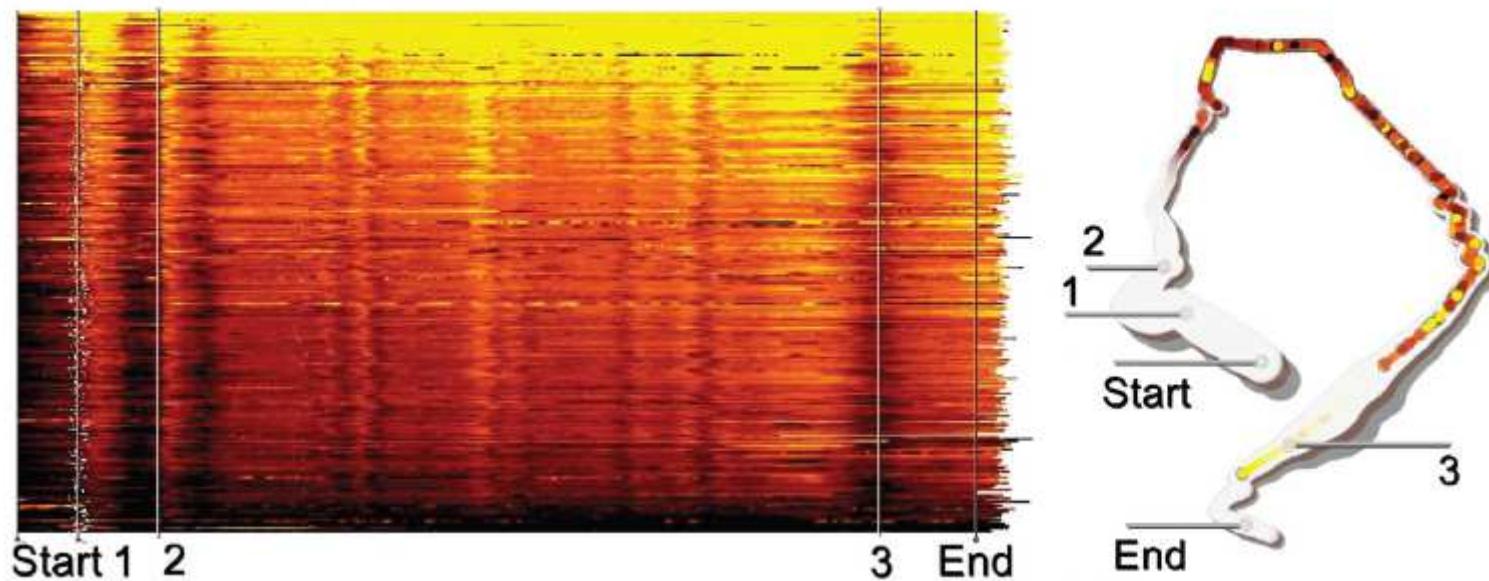


Figure 10 – Exemple de « Heatmap » sur un client SMB de PG&E : comprendre la consommation et les besoins potentiels du client

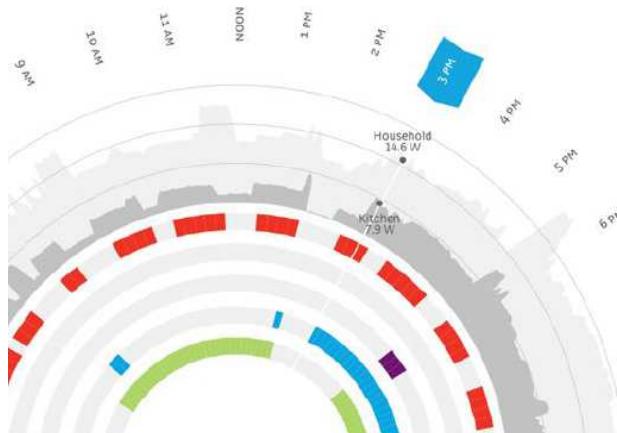
Visualisation de séries temporelles

<https://rafaeltorchelsen.wordpress.com/category/publications/>

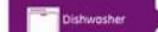
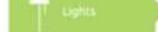
- Marking alignments of the linear heatmap in the course.



Visualisation de séries temporelles

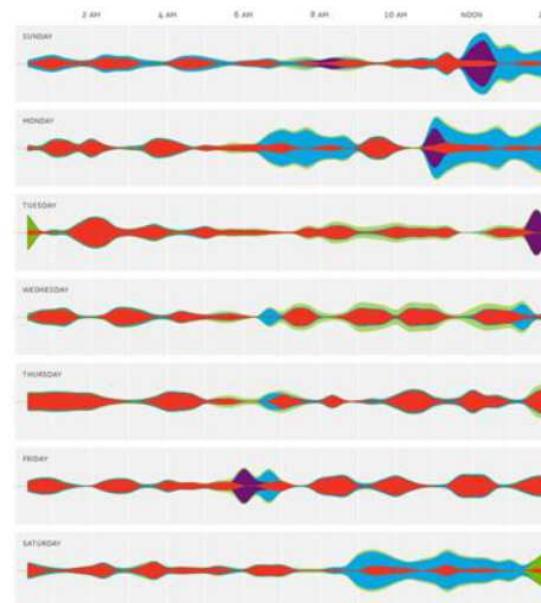


Tap on an appliance to see how much it cost to run over the course of a week.

-  Refrigerator
-  Range
-  Dishwasher
-  Plugs
-  Lights

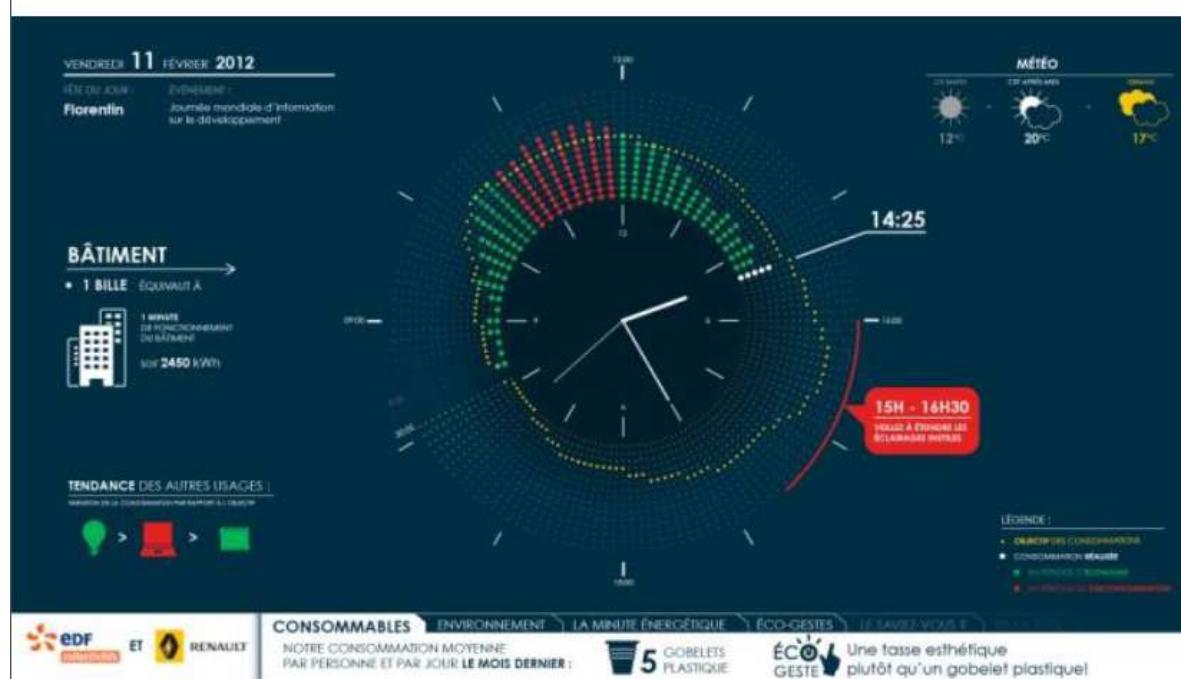
Powering the kitchen,
Fathom, Ben Fry en collaboration avec GE,
2011

The data utilized in this application comes from
a sample household equipped with Brillion
technology over one month's time



<http://visualization.geblogs.com/visualization/kitchen/>

Visualisation de séries temporelles

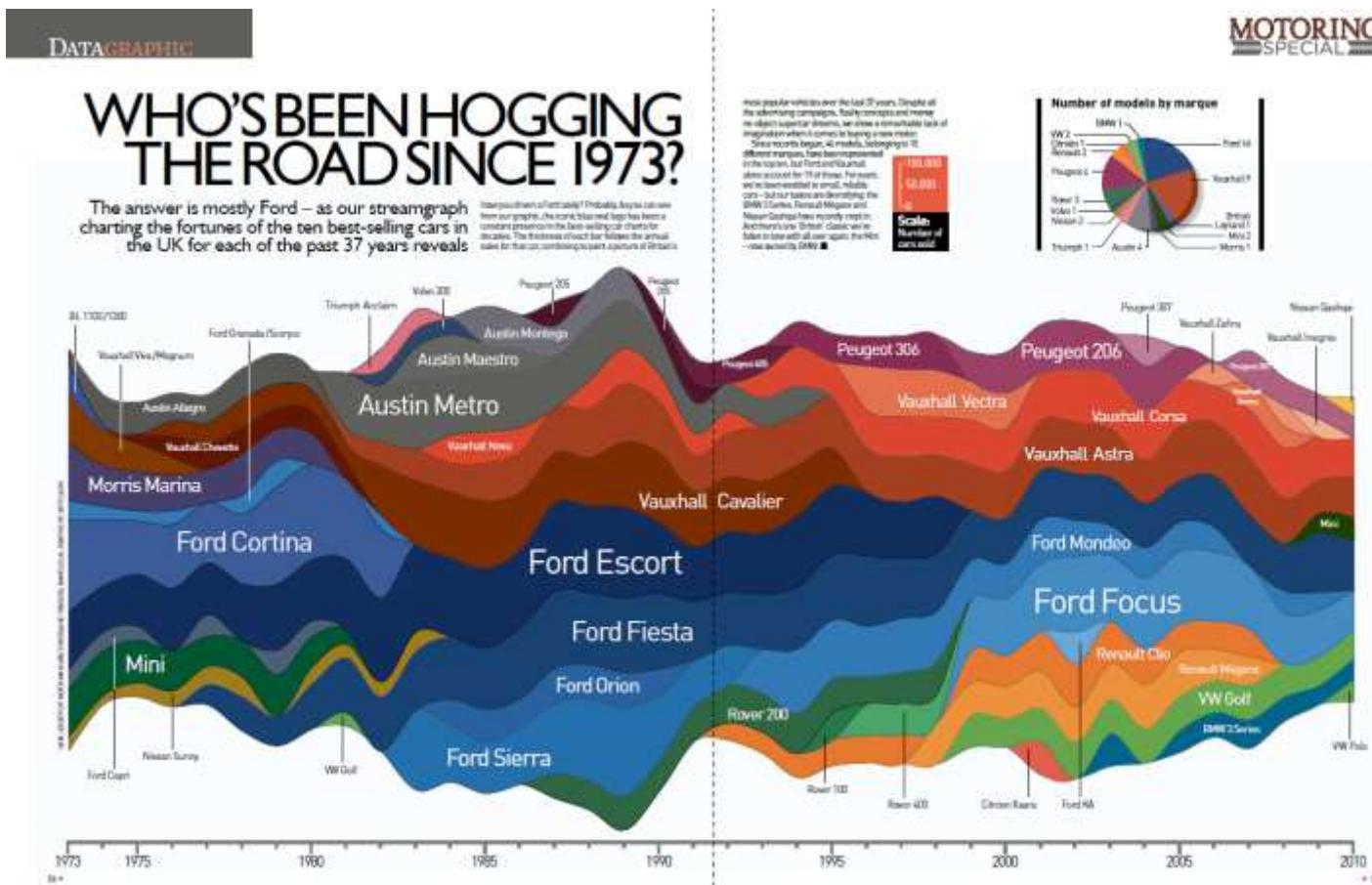


**Horloge Energétique,
EDF, 2013**

Outil qui permet de mesurer et visualiser la consommation énergétique d'un bâtiment en temps réel, et vise à faire évoluer les comportements de chacun par une prise de conscience de l'impact de ses propres gestes sur la consommation énergétique globale...

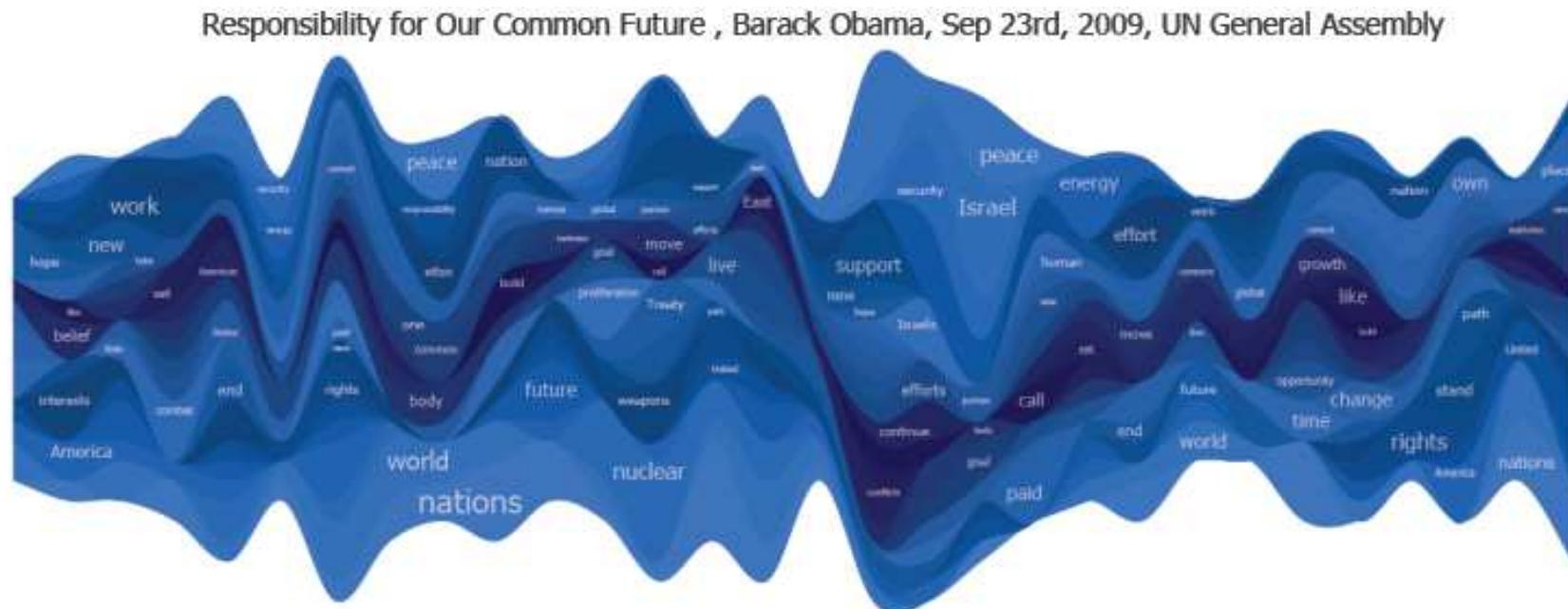
Visualisation de séries temporelles

<http://www.neoformix.com>



Visualisation de séries temporelles

<http://www.neoformix.com>



Analyse exploratoire de séries temporelles

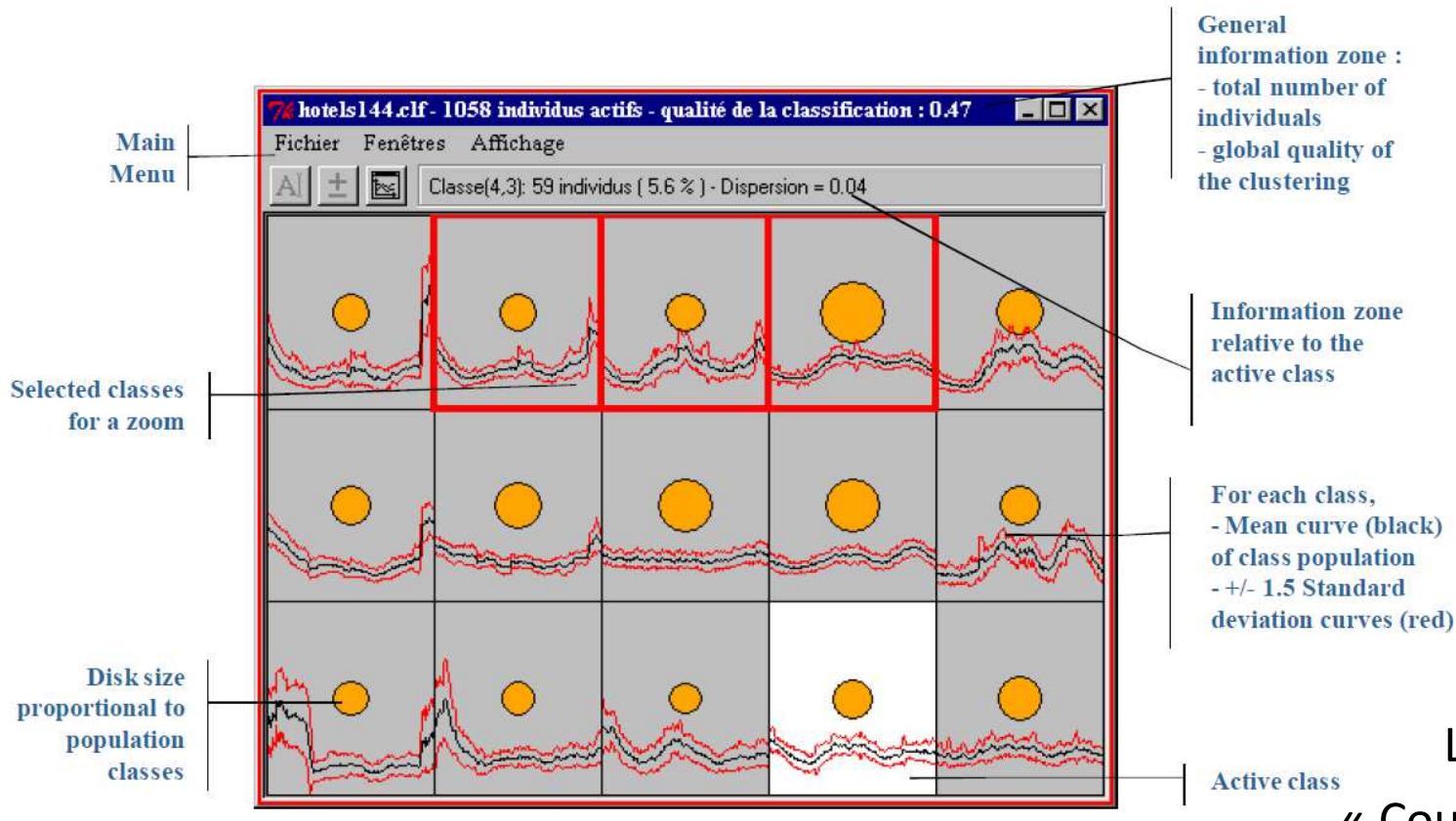
- Requêtage de séries temporelles
(Similarity based time series retrieval)
- Visualisation
- Classification automatique (clustering)
- Recherche de motifs fréquents

Classification automatique (clustering)

- Clustering
 - Série mono-dimensionnelle
 - Typologie de périodes: ex. consommation journalière
 - 1 individu = 1 jour
 - Série multi-dimensionnelle
 - Typologie d'états: ex. capteurs sur un matériel
 - 1 individu = 1 instant

Clustering de périodes

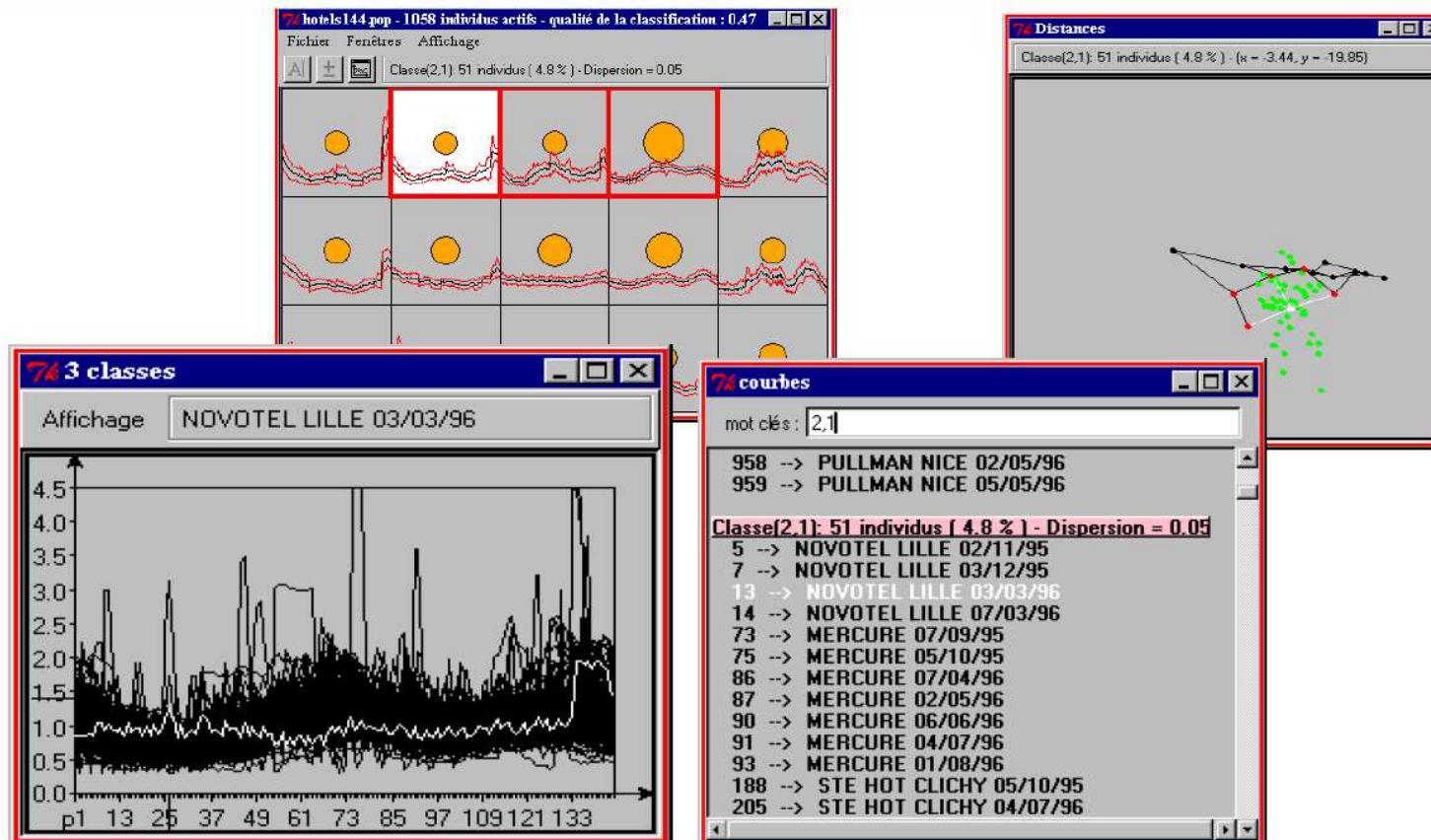
Cartes auto-organisatrices de Kohonen



Logiciel EDF
« Courboscope »

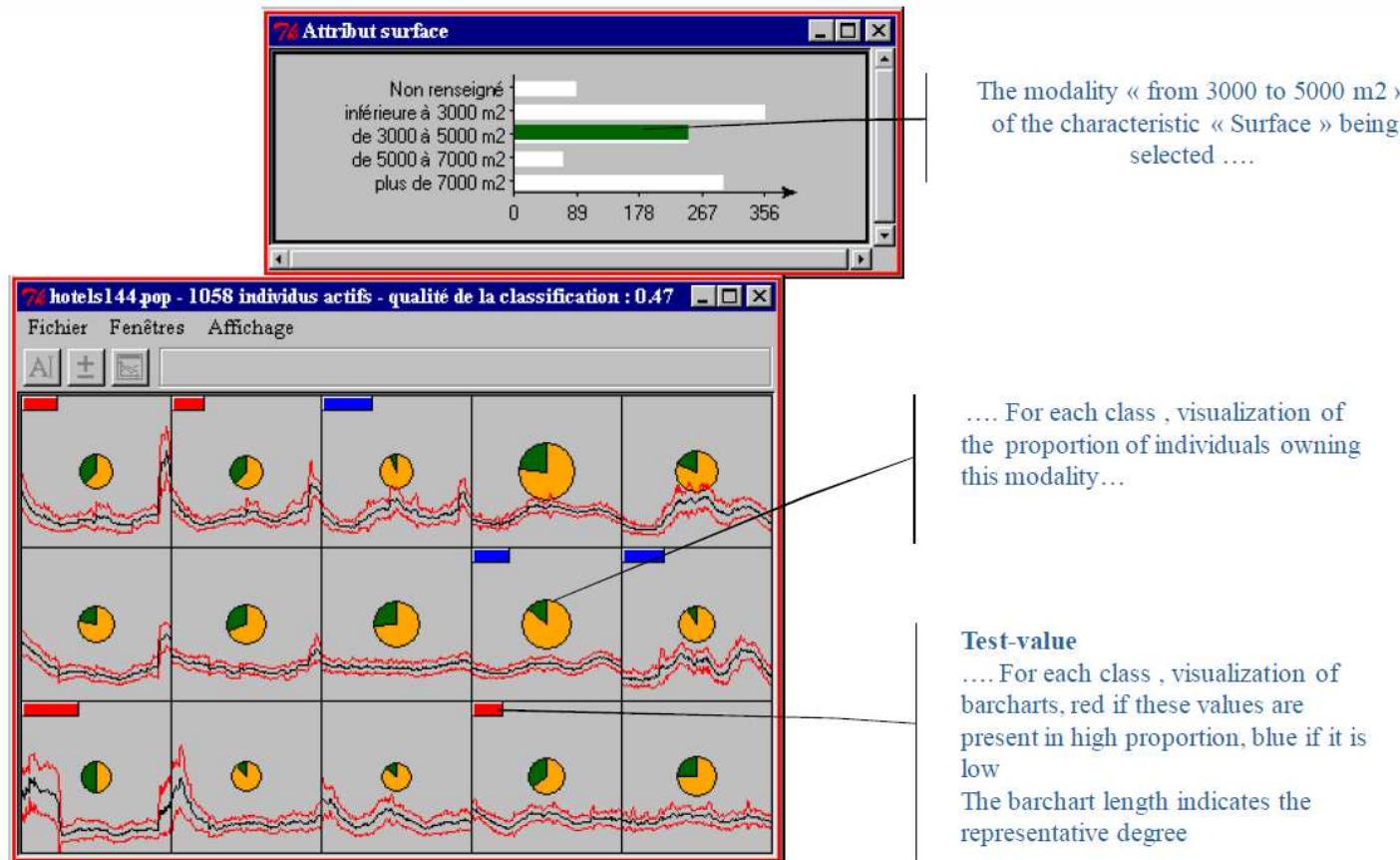
Clustering de périodes

Interprétation des classes



Clustering de périodes

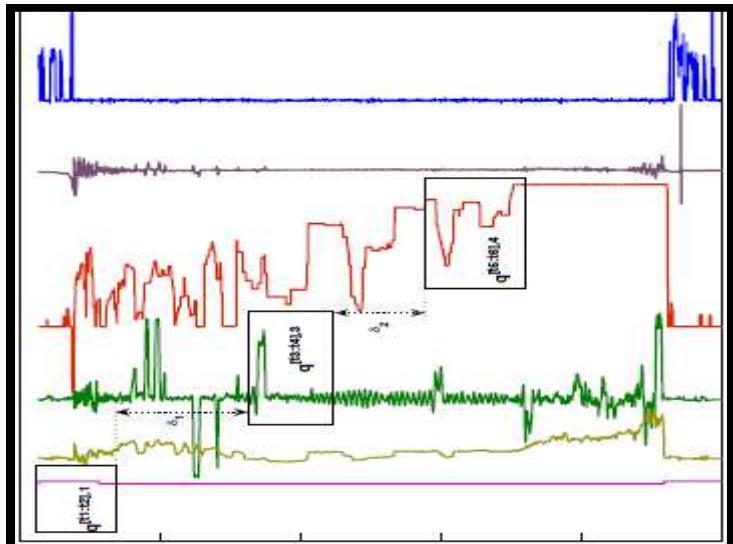
Interprétation des classes



Clustering d'états

Série multi-dimensionnelle

- Plusieurs capteurs sur un matériel

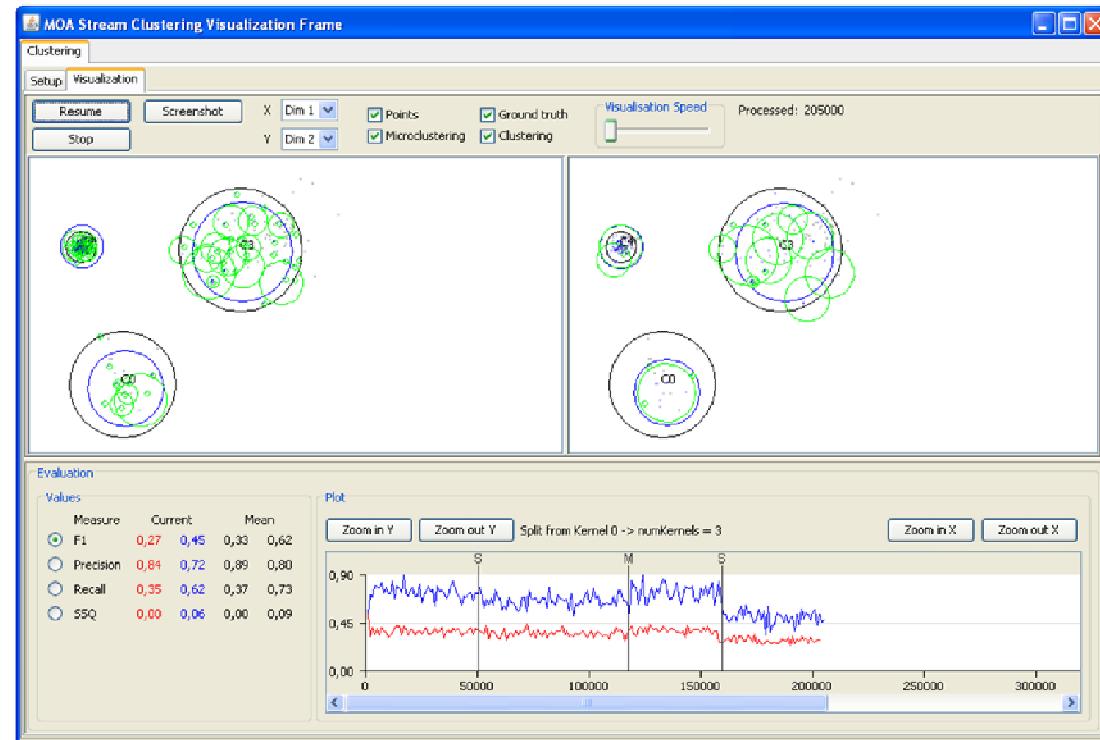


t	C1	C2	...	C6
1	3,27	18,12	...	192
2	3,28	19,5	...	198
...
...
...
...
n	3,32	26,3	...	167

Clustering d'états

Surveillance d'un matériel

- 1 instant = 1 individu
- Classes d'états : états sains / états en anomalie
- Clustering « batch » / clustering « évolutif »



Source :

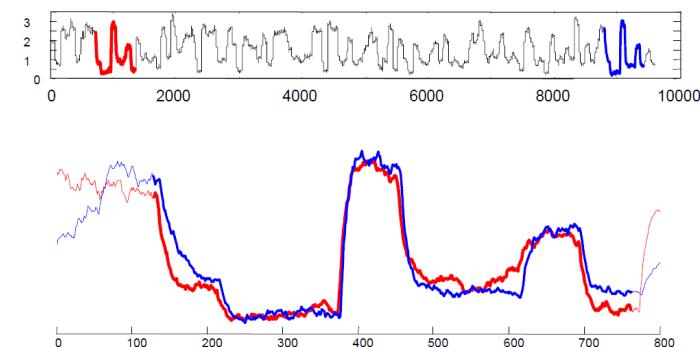
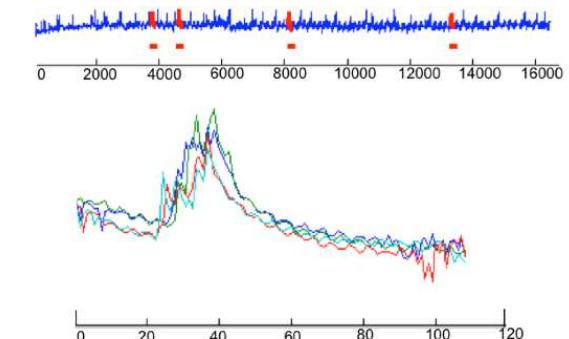
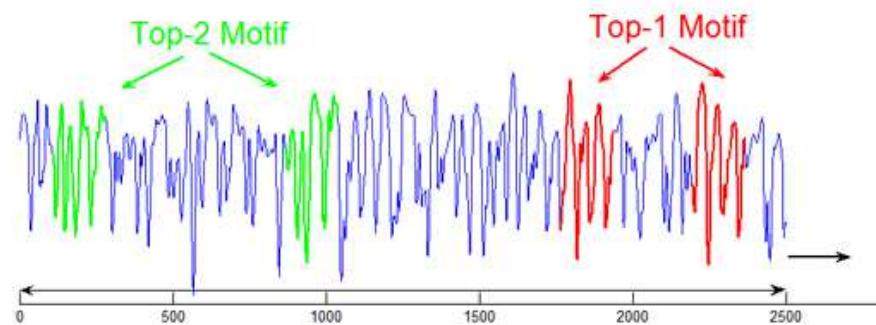
<http://moa.cms.waikato.ac.nz/details/stream-clustering/>

Analyse exploratoire de séries temporelles

- Requêtage de séries temporelles
(Similarity based time series retrieval)
- Visualisation
- Classification automatique (clustering)
- Recherche de motifs fréquents

Recherche de motifs fréquents

- Dans une ou plusieurs séries
- Caractère contigu / non contigu du motif
- Longueur du motif fixe / variable
- Algorithmes
 - Calcul de distance sur fenêtre glissante
 - Transformation en séquence de symboles
 - En général coûteux



Recherche de motifs fréquents

- Exemple

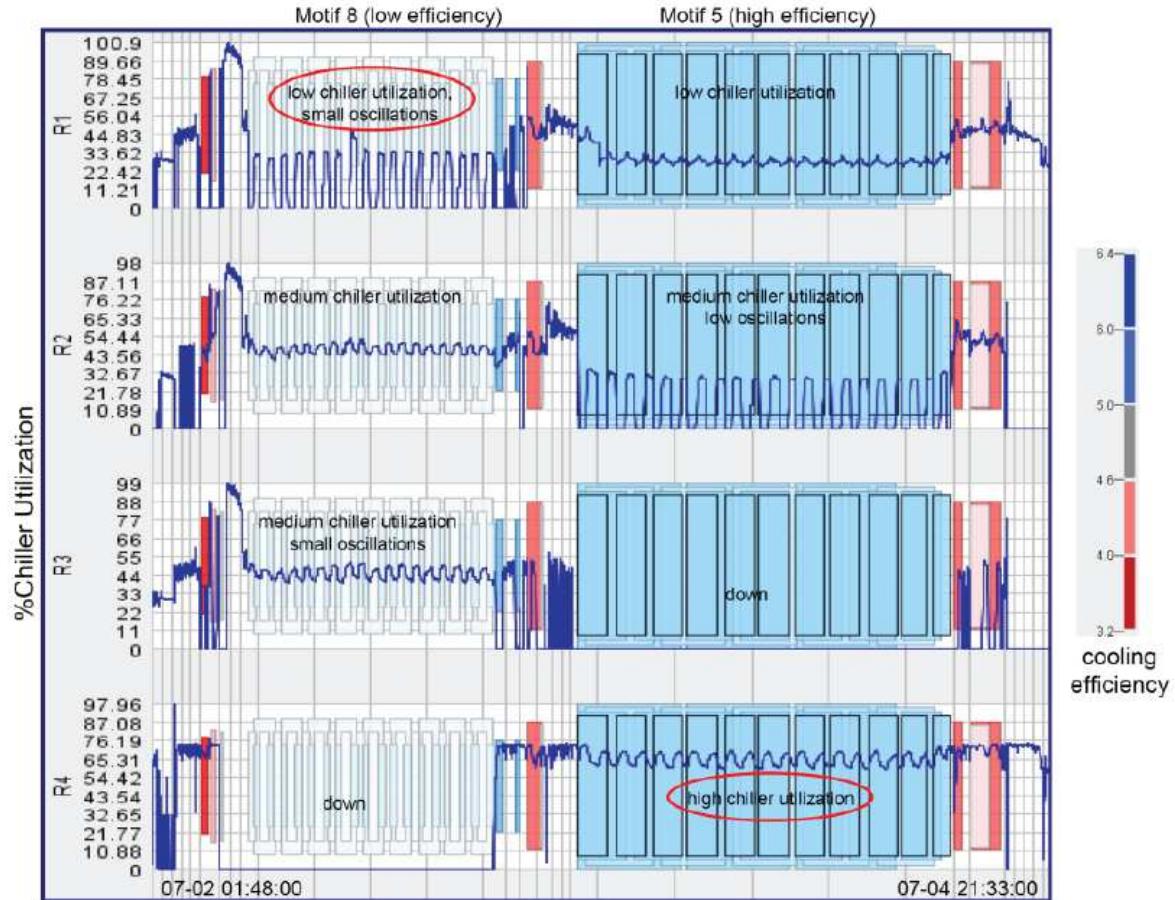


Figure 8: Motifs 5 and 8 are Enlarged to Compare their Chiller Utilization
 Motif 5 is more efficient than motif 8. Motif 8's chillers R1 and R3 have some oscillatory behavior.
 (x-axis: 07-02 01:48 to 07-04 21:33, y-axis: %utilization of chillers R1-R4, color: cooling efficiency).

Source : "Visual exploration of Frequent Patterns in Large Multivariate Time Series", M. Hao, M. Marwah, H. Janetzko, R. Sharma, D. A. Keim, U. Dayal, D. Patnaik, N. Ramakrishna, *Information Visualization*, 0(0) 1–13, 2011.

Analyse des séries temporelles

- Data analytics
 - Exploratory analysis / unsupervised learning
 - Explorer, synthétiser des données
 - Predictive analysis / supervised learning
 - Utiliser des données pour prédire/prévoir

Analyse prédictive de séries temporelles

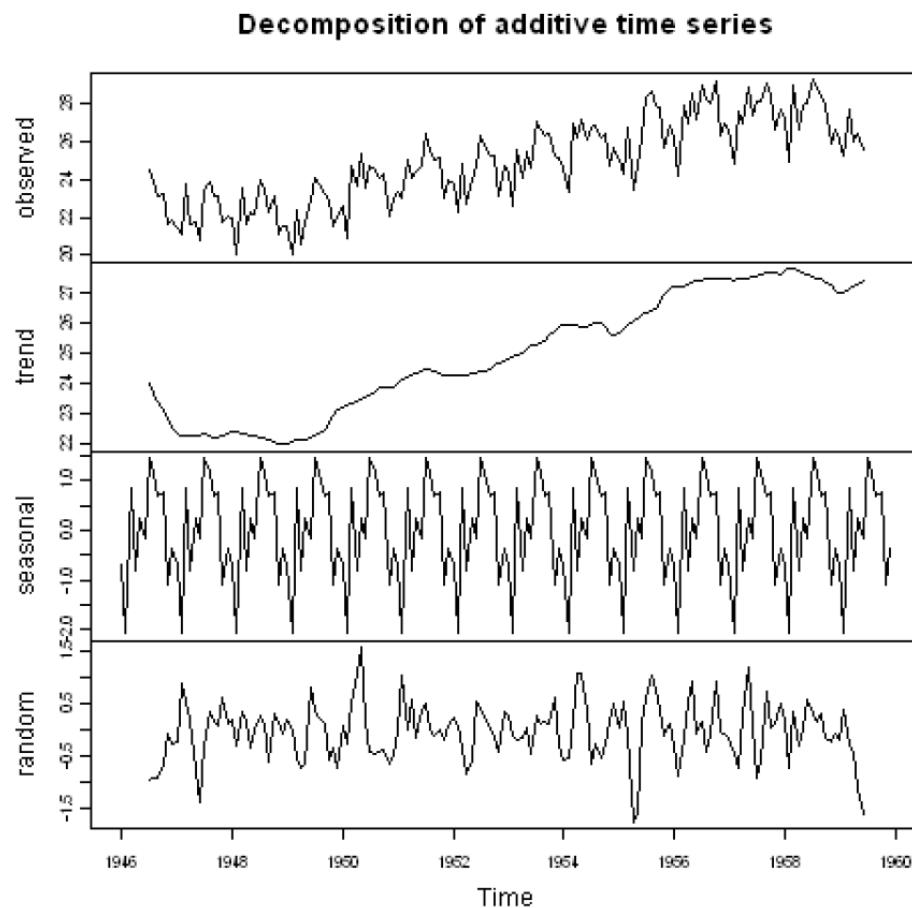
- Prévision de séries temporelles
- Détection d'anomalies

Méthodes de prévision

- Préparation des données
 - Données manquantes, outliers, ...
 - Mise à un pas régulier
- Analyse exploratoire
 - Tendance, saisonnalité, ...
 - Autocorrélation temporelle
- Choix d'un modèle de prévision
- Evaluation de la qualité du modèle

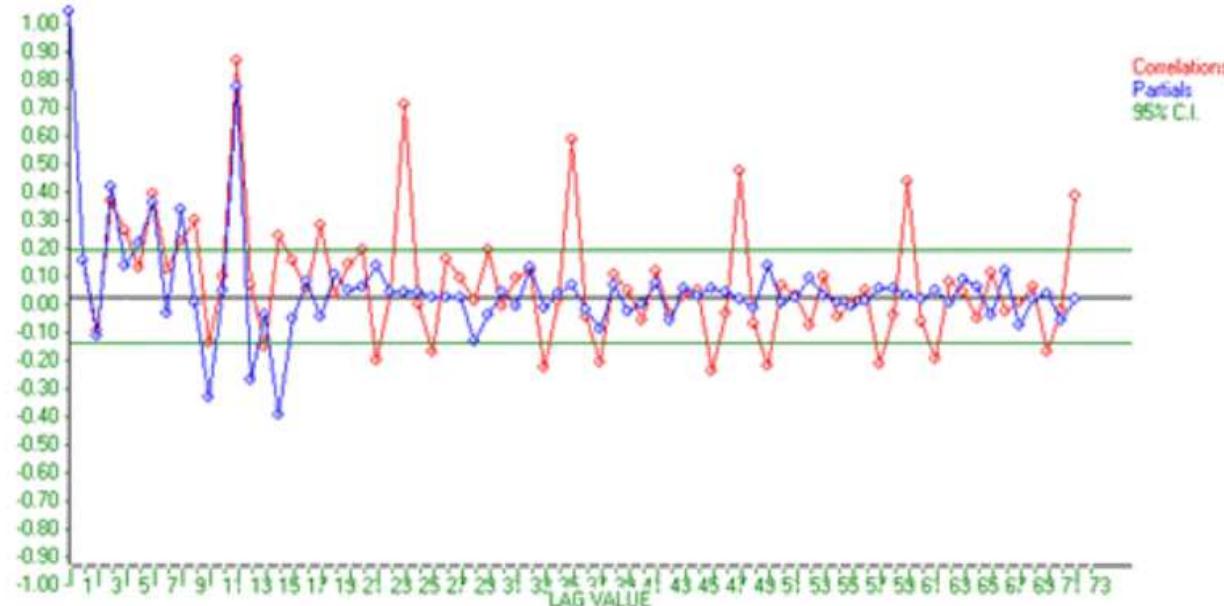
Méthodes de prévision

Stationnarité, décomposition, tendance, saisonnalité, ...



Méthodes de prévision

Autocorrélation temporelle



Méthodes de prévision

Choix d'un modèle de prévision

- Très nombreux modèles de prévision, dépendant des caractéristiques de la série, de la présence de prédicteurs « exogènes », ...
- Lissage exponentiel, Holt, Winters
- Modèles auto-régressifs AR, MA, ARMA, ARIMA
- Ajouts de variables explicatives, ex. ARIMAX, GAM

Méthodes de prévision

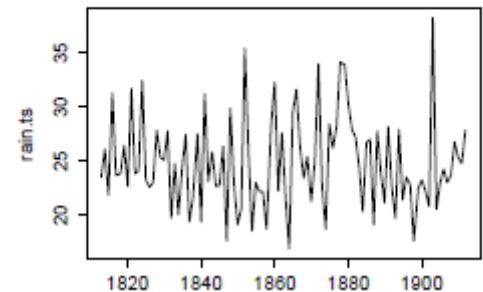
Exemples

- Lissage exponentiel simple
 - pas de tendance, pas de saisonnalité, court terme

$$\hat{x}_T(1) = c_0 x_T + c_1 x_{T-1} + c_2 x_{T-2} + \dots$$

$$c_i = \alpha (1 - \alpha)^i$$

Annual Rainfall in London



- Modèles GAM (Generalized Additive Models)

$$y_t = f_1(x_t^1) + f_2(x_t^2) + \dots + f(x_t^3, x_t^4) + \dots + \varepsilon_t$$

$$\min_{\beta, f_j} \|y - f_1(x_1) - f_2(x_2) - \dots\|^2 + \lambda_1 \int f_1''(x)^2 dx + \lambda_2 \int f_2''(x)^2 dx + \dots$$

Méthodes de prévision

Evaluation de la qualité du modèle

- Série d'apprentissage, série de test

• Mean Error

$$ME = \frac{1}{n} \sum_{t=1}^n (x_t - \hat{x}_t)$$

• Mean Percentage Error

$$MPE = \frac{1}{n} \sum_{t=1}^n \left(\frac{x_t - \hat{x}_t}{x_t} \right)$$

• Mean Absolute Error

$$MAE = \frac{1}{n} \sum_{t=1}^n |x_t - \hat{x}_t|$$

• Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (x_t - \hat{x}_t)^2}$$

• Mean Absolute Percentage Error

$$MAPE = \frac{100}{n} \sum_{t=1}^n \left| \frac{x_t - \hat{x}_t}{x_t} \right|$$

Analyse prédictive de séries temporelles

- Prévision de séries temporelles
- Détection d'anomalies

Détection d'anomalies

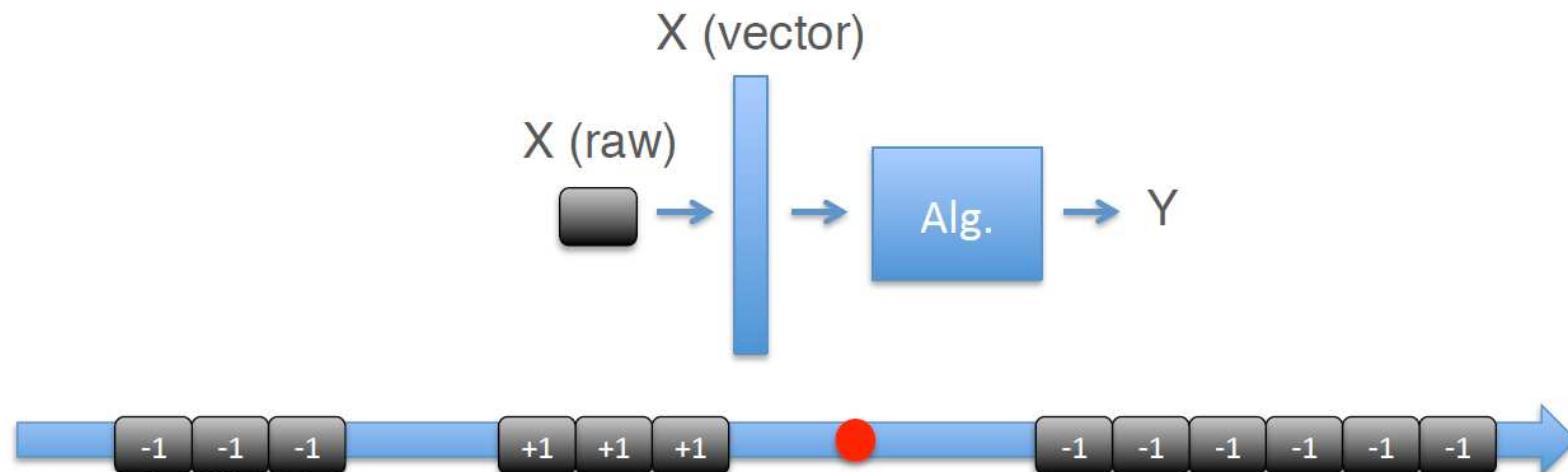
- Par apprentissage à partir de données versus à l'aide de modèles physiques / règles d'expert
- Maintenance planifiée versus prédictive (condition-based maintenance)

Détection d'anomalies

- Se ramener à l'apprentissage d'un modèle prédictif classique
- Prédicteurs
 - Données statiques :
 - Liées à l'équipement/composant/capteur : ID, modèle, année, constructeur, spécifications, ...
 - Liées à l'utilisateur/client : ID, type, contrat, localisation, ...
 - Données dynamiques
 - Données de maintenance
 - Données continues
 - Données qualitatives
- Variable à prédire
 - Anomalies : apparition de l'événement, délai de l'apparition
- Constitution d'un ensemble d'apprentissage
 - Pannes réelles, annotation à dire d'expert (ex. alarmes)

Détection d'anomalies

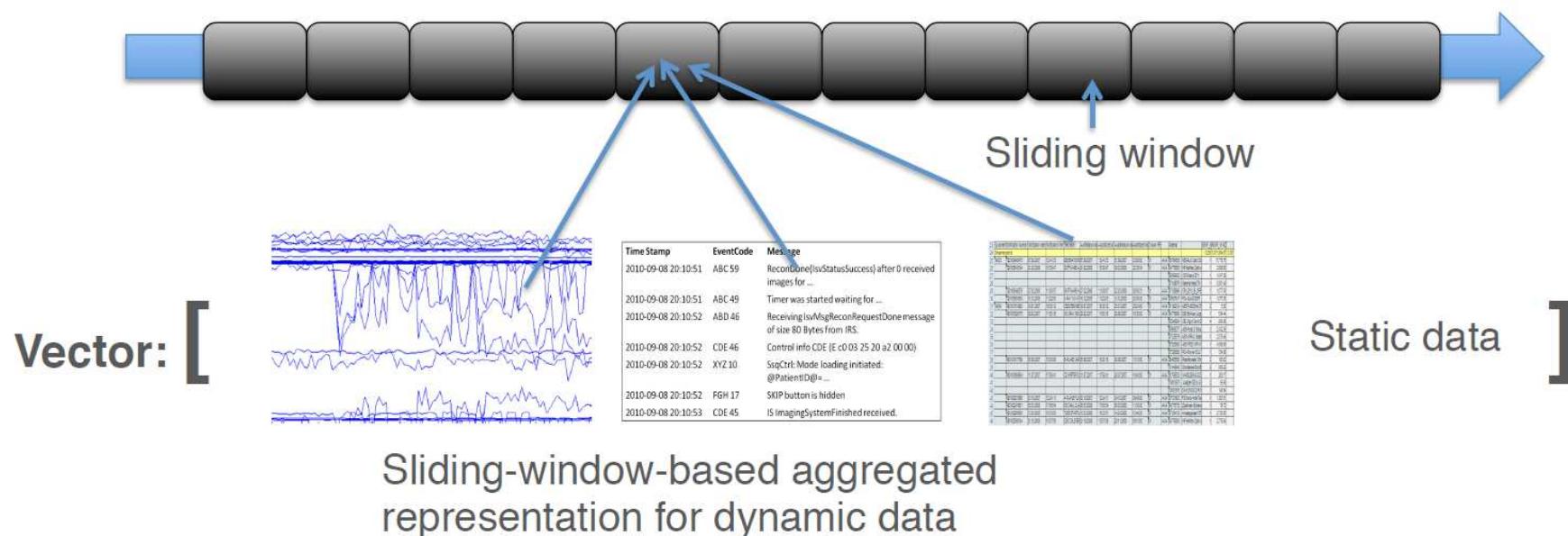
- Nombreux modèles
- Exemple simple
 - Prédiction de l'apparition d'une anomalie : binaire
 - Prédire suffisamment tôt



Source : "Predictive Maintenance in a Machine Learning Perspective", IEEE BigData 2015 Tutorial", Zhuang (John) Wang.

Détection d'anomalies

- Construction des prédicteurs (« feature engineering »)
- Découpage en fenêtres temporelles (très lié à l'application)
- Sliding window based feature representation for disparate data types



Détection d'anomalies

Featurizing Continuous Time Series (cont.)

- Within each window, for each uni-variate time series, extract:
 - Aggregate basic characteristics
 - Mean, median, std, max, min, #observations, slope, #observations beyond normal range, zero crossing rate, last N value(s), ...
 - Aggregate temporal characteristics
 - Feature value differencing based on time/window
 - # sequences (pre-defined or learned by sequential mining, see Moerchen, 10')
- Then concatenate them as a vector to represent this sliding window

Source : "Predictive Maintenance in a Machine Learning Perspective", IEEE BigData 2015 Tutorial", Zhuang (John) Wang.

Détection d'anomalies

Featurizing Categorical Time Series

- Using bag-of-objects representation:
 - Bag = sliding window
 - Object =
 - Event/severity code: e.g. ABC 123, XYZ 456, ERROR/INFO/ DEBUG, ...
 - Text message template
 - Keyword: list of domain-specific keywords, e.g. leak, break, burn, ...

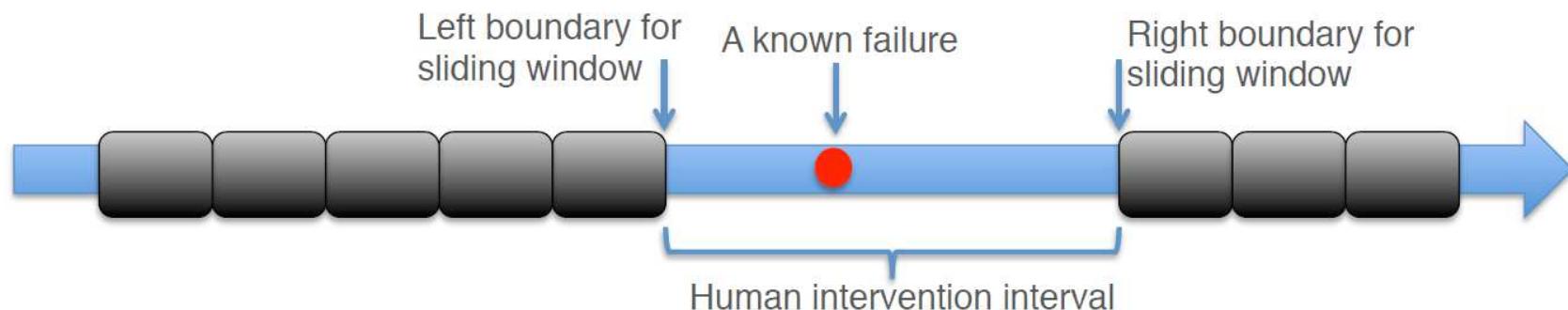
Featurizing Service Data and Static Data

- Service data is usually coarsely time-stamped. Bag-of-objects representation can be used to extract domain-interested information, such as:
 - # services in past N months
 - Average/total service time/interval
 - Occurrence of certain (combination of) service(s) in the past
 - Average/total service cost
 - Keywords
 - ...
- Static data (such as model, year, manufactory, location, customer info.) which are equipment/machine specific can be featurized by standard feature engineering techniques

Détection d'anomalies

Challenge – How to Choose Sliding Window?

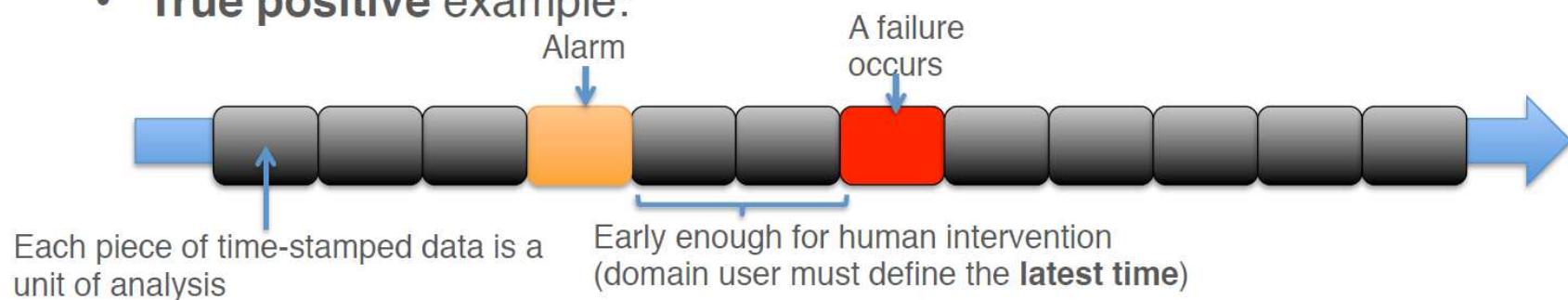
- The **width** of sliding window should be first coarsely chosen based on domain knowledge and then fine tuned as a hyper-parameter based on evaluation
- The **increment** of sliding window depends on prediction frequency (i.e. how often the learned model is used to score new data)
- The **position** of sliding window depends on the domain-specific predictive maintenance process
 - What is the latest time for human intervention?
 - How long is human intervention?



Détection d'anomalies

- Evaluation de la qualité du modèle
 - Prédire « au bon moment » / trop tard/tôt

- **True positive example:**



- **False positive example:**



Détection d'anomalies

- Evaluation de la qualité du modèle
 - Prise en compte de coûts de mauvais classement

TP = \$50	FN = -\$10
FP = -\$100	TN = \$1

Confusion matrix
 (for each model-threshold pair)

	P' (Predicted)	N' (Predicted)
P (Actual)	True Positive	False Negative
N (Actual)	False Positive	True Negative

Détection d'anomalies

- Evaluation de la qualité du modèle
 - Ensemble d'apprentissage / ensemble test
 - Même matériel / transfert à un autre matériel
 - Evénements rares

AGENDA

- Exemples de motivation
- Données brutes de capteurs
- Préparation des données
- Similarités entre séries temporelles
- Analyse de séries temporelles
 - Analyse exploratoire
 - Analyse prédictive
- **Etude de cas**

Références

- “Data mining: the textbook”, C.C.Aggarwal, Springer 2015.
- “Mining time series”, CS240B Notes by Carlo Zaniolo, UCLA, Computer Science Department, 2015.
- “Functional data analysis”, Ramsay & Silverman, Springer 1997.
- “Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining”, C.Cassisi, P.Montalto, M.Aliotta, A.Cannata and A.Pulvirenti, Advances in Data Mining Knowledge Discovery and Applications, Associate Prof. Adem Karahoca (Ed.), InTech, DOI: 10.5772/49941, (2012).
- “Searching time series with Hadoop in an electric power company”, Bérard A., Hébrail G., Workshop BIG'MINE 2013, ACM-KDD Conference, Chicago, 2013.
- “A Tour Through the Visualization Zoo”, By Jeffrey Heer, Michael Bostock, Vadim Ogievetsky, Communications of the ACM, Vol. 53 No. 6, Pages 59-67, 10.1145/1743546.1743567
- “Visual exploration of Frequent Patterns in Large Multivariate Time Series”, M. Hao, M. Marwah, H. Janetzko, R. Sharma, D. A. Keim, U. Dayal, D. Patnaik, N. Ramakrishna, Information Visualization, 0(0) 1–13, 2011.
- “Introductory Time Series with R”, Paul S.P. Cowpertwait, 2006 SPRINGER.
- “Time Series Analysis in a Nutshell using R”, J.J.M. Rijpkema, Eindhoven University of Technology, dept. Mathematics & Computer Science, 2012.
- “Automatic Time Series Forecasting: The forecast Package for R”, Rob J. Hyndman Monash University Yeasmin Khandakar Monash University, Journal of Statistical, Volume 27, Issue , 2008.
- “Predictive Maintenance in a Machine Learning Perspective”, IEEE BigData 2015 Tutorial, Zhuang (John) Wang.