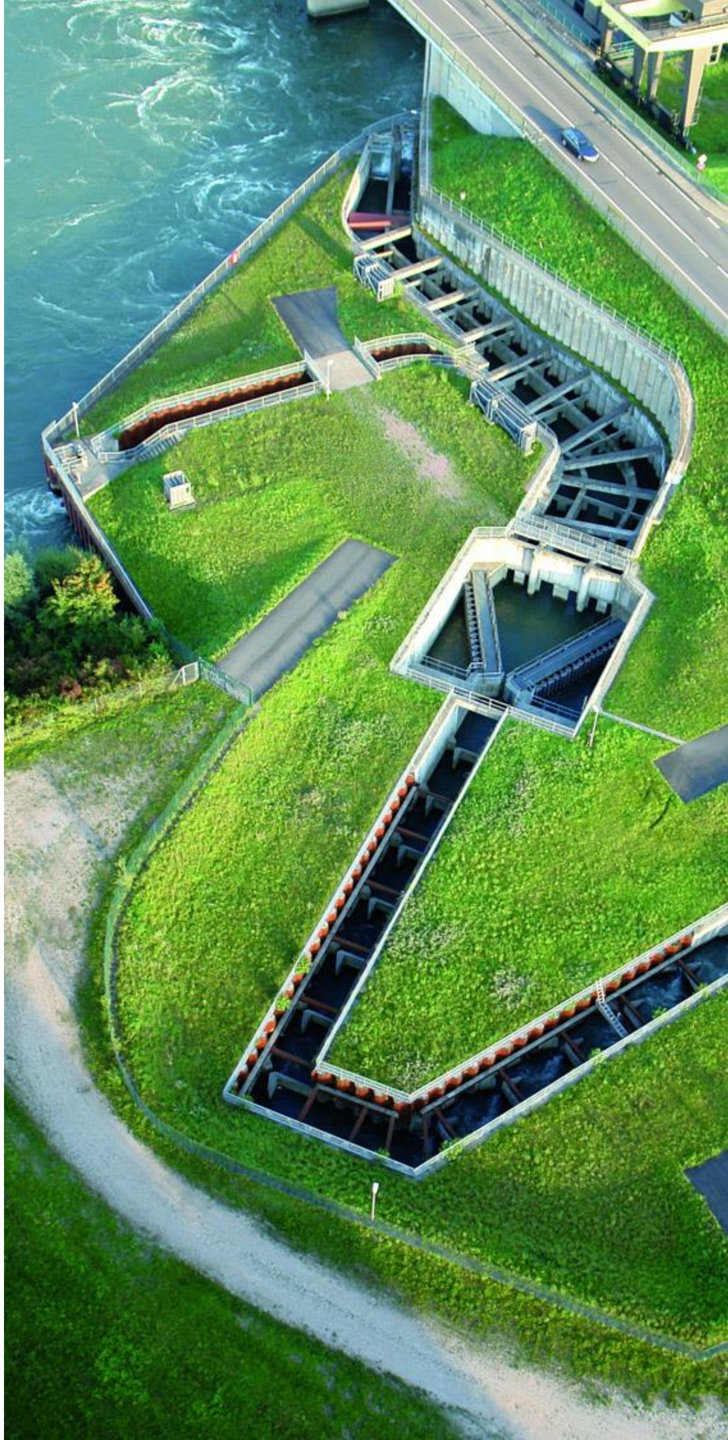
PRIVACY-PRESERVING USE OF INDIVIDUAL SMART METERING DATA FOR CUSTOMER SERVICES

Georges Hébrail, EDF R&D, France

Joint work with T.Allard, F.Masseglia, E.Pacitti
INRIA Zenith and University of Montpellier, France

2nd International Interdisciplinary Workshop
on Time Series Analysis
CNRS/MASTODONS

December 13th, 2016
University Paris Descartes




Smart Grid
... We're Watching


Linky
GRDF
BIG LINKY
is WATCHING
YOU!
UFC-Que Choisir
Biz

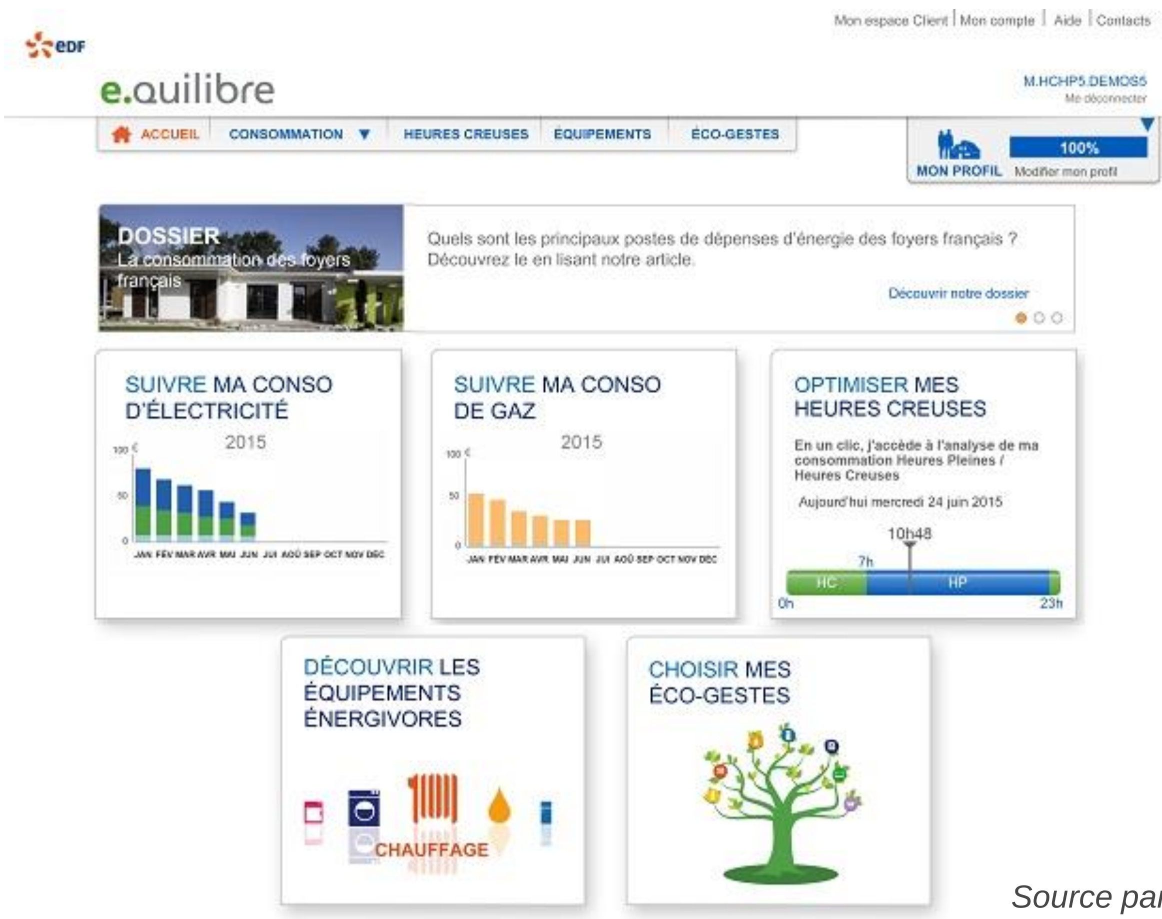
SMART METERS AND CONNECTED OBJECTS

- **Deployment of smart meters** (Linky project in France)
  - From 2016 to 2020 (35M meters)
  - Remote turning power on/off, remote readings and billing
  - Readings up to every 10 minutes to the supplier/distributor
  - Readings up to 2s on premisses

- **Deployment of connected objects in households** ('smart home')

NEW SERVICES TO CUSTOMERS

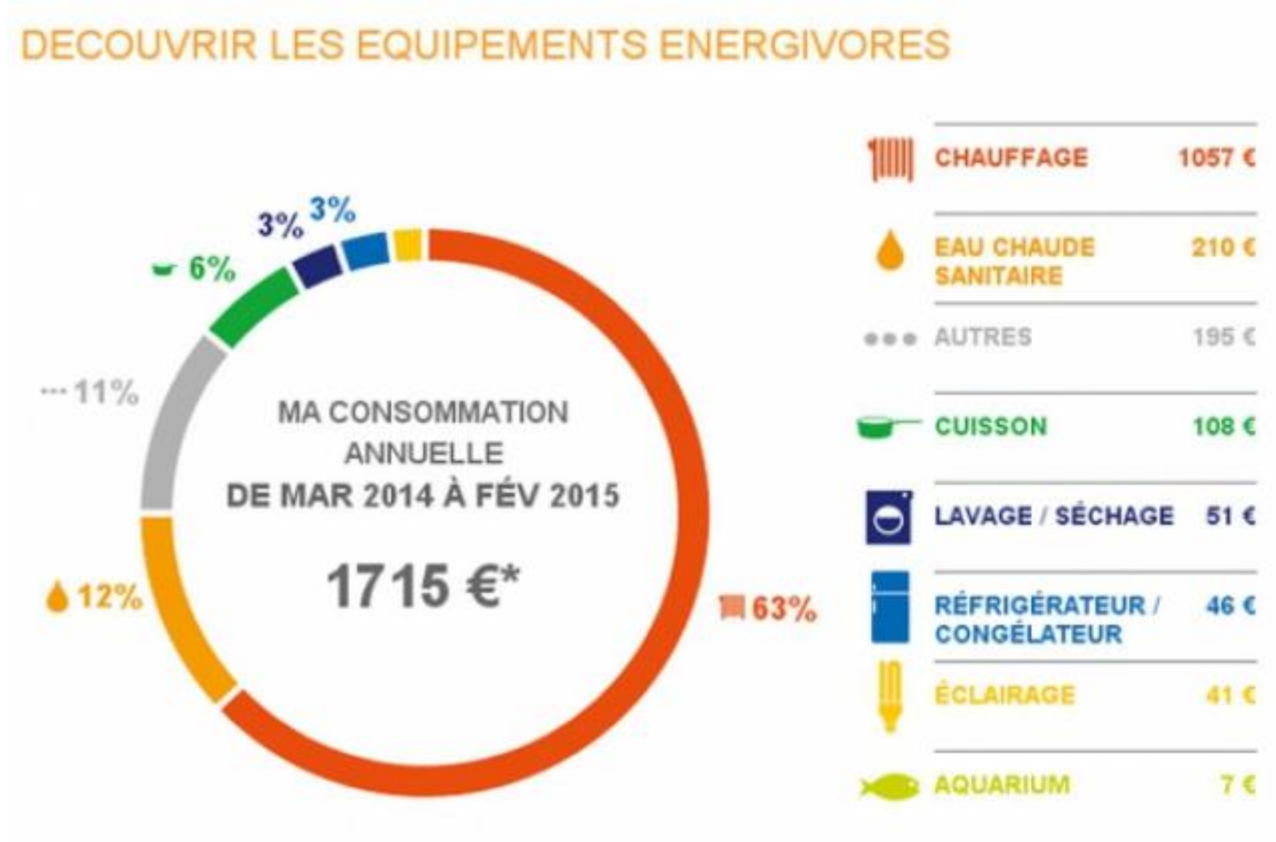
- **Using smart meter readings for energy efficiency diagnosis and advice**



Source particulier.edf.fr

NEW SERVICES TO CUSTOMERS

- **Using smart meter readings for energy efficiency diagnosis and advice**



Source particulier.edf.fr

NEW SERVICES TO CUSTOMERS

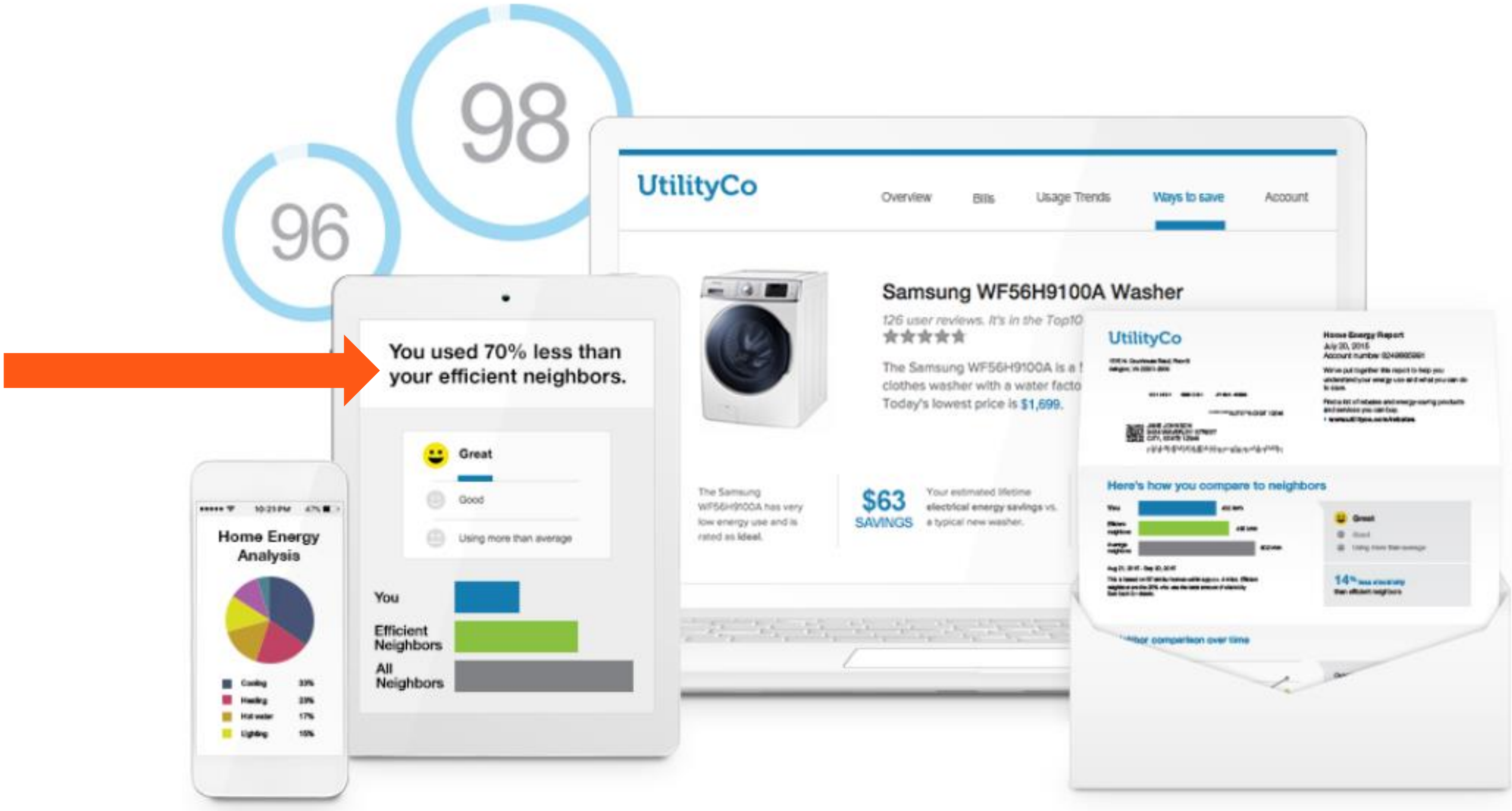
- **Using smart meter readings for energy efficiency diagnosis and advice**

Green Button Download My Data ®

Green Button Connect My Data

NEW SERVICES TO CUSTOMERS

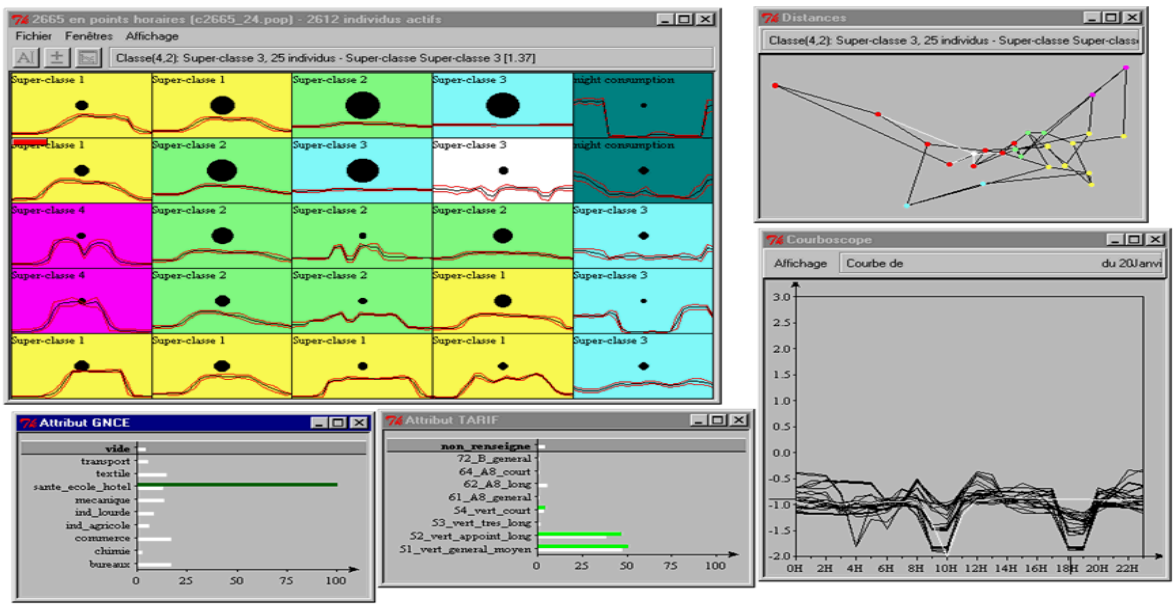
- **Using smart meter readings for energy efficiency diagnosis and advice**



Source www.opower.com

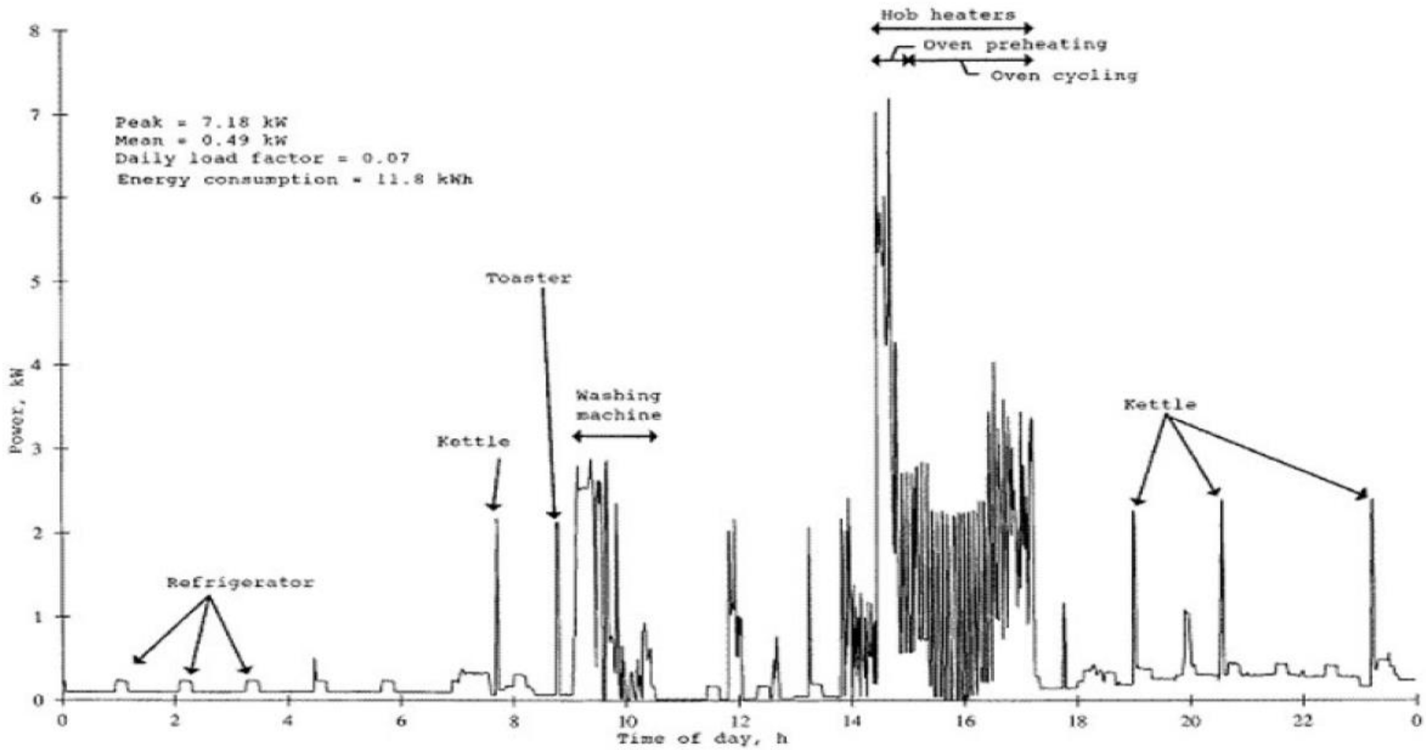
NEW SERVICES TO CUSTOMERS

- **Using smart meter readings for energy efficiency diagnosis and advice**
  - One standard approach: comparison to « neighbors »
    - Storage of individual consumption curves in a centralized data warehouse
    - Construction of (daily/weekly) profiles by clustering of individual curves
    - Association of house/equipment/occupants characteristics to clusters
    - Comparison of individual data with profiles



GREAT … BUT …

- **Consumption data becomes more sensitive at a higher sampling rate**
  - Presence/absence, number of people in the house
  - Human activity (cooking, shower, TV, …)



Household electrical consumption example

Newborough et P. Augood, « Demand-side management opportunities for the UK domestic sector », Generation, Transmission and Distribution, IEE Proceedings-, vol. 146, n° 3, p. 283 -293, mai 1999.

PRIVACY-PRESERVING SERVICES TO CUSTOMERS

Do the same job but with privacy preservation of individual electric power consumption curve !

→ « Chiaroscuro »

- **Basic idea**
  - Customer advice is computed locally (can easily be private)
  - Construction of profiles with associated household characteristics

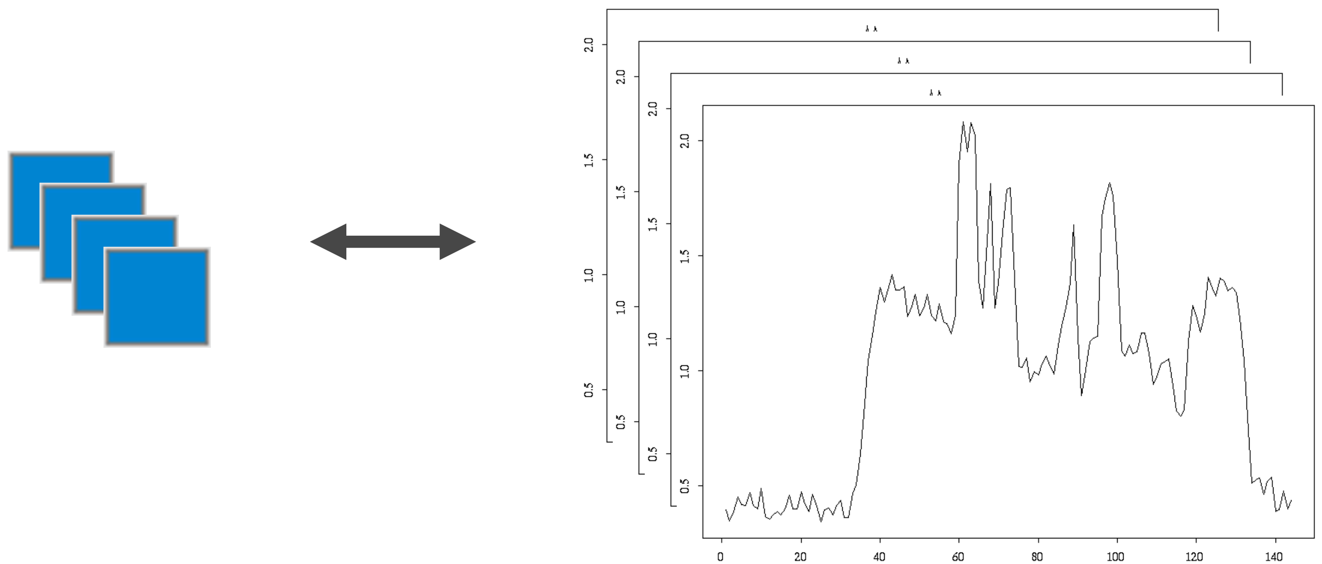
  → New approach of privacy-preserving clustering of individual consumption curves

PRIVACY-PRESERVING TIME SERIES CLUSTERING

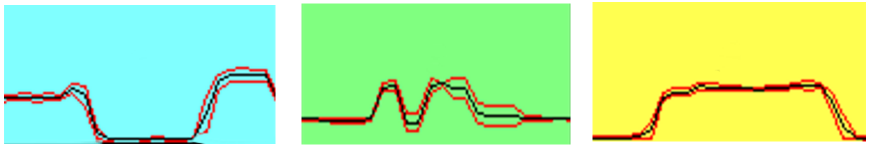
- **Privacy-preserving distributed clustering**
- **P2P infrastructure**
- **Evaluation**

PRIVACY-PRESERVING DISTRIBUTED CLUSTERING

- **Data input**
  - N geographically distributed individual daily electric power consumption time series
  - 24 dimensions vectors if hourly data, 144 dimensions data if 10’ data
  - Euclidian distance on (normalized) coordinates

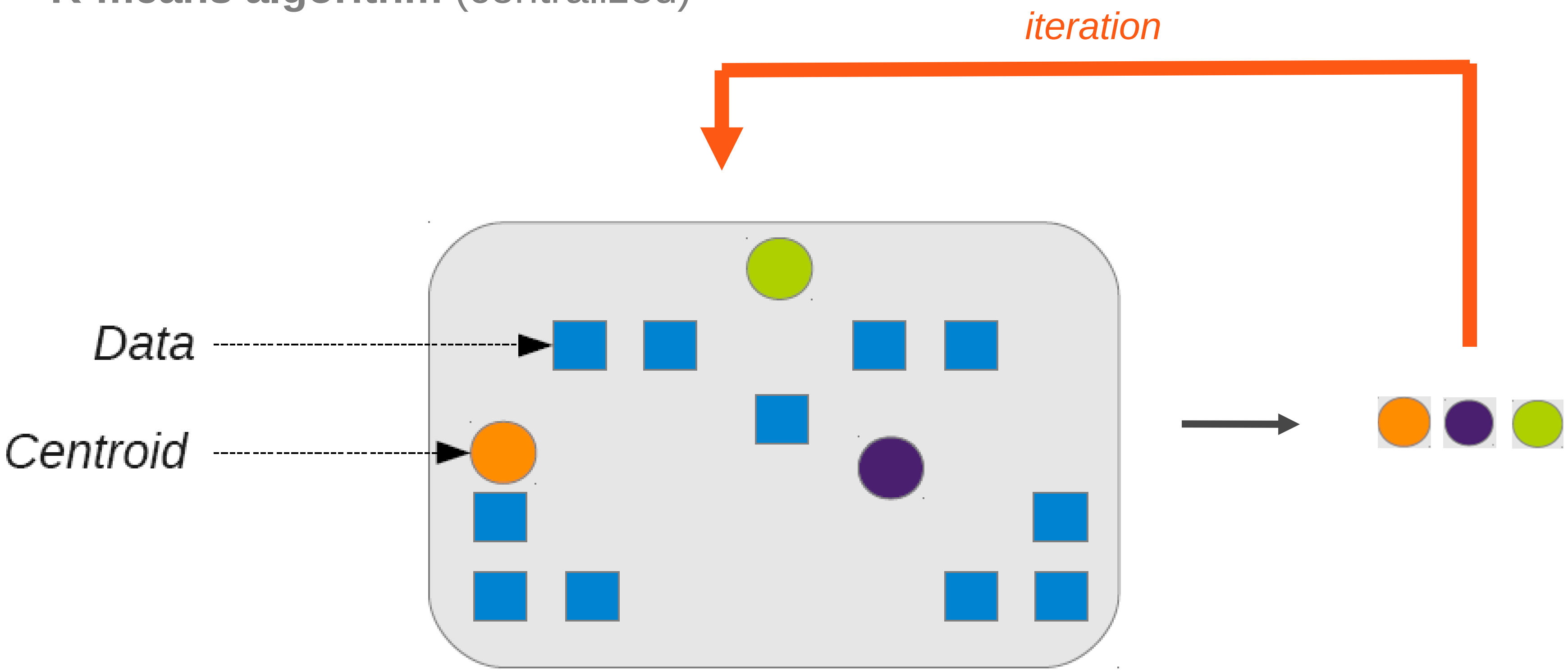


- **Output result**
  - K time-series profiles (24 ou 144 dimensions)



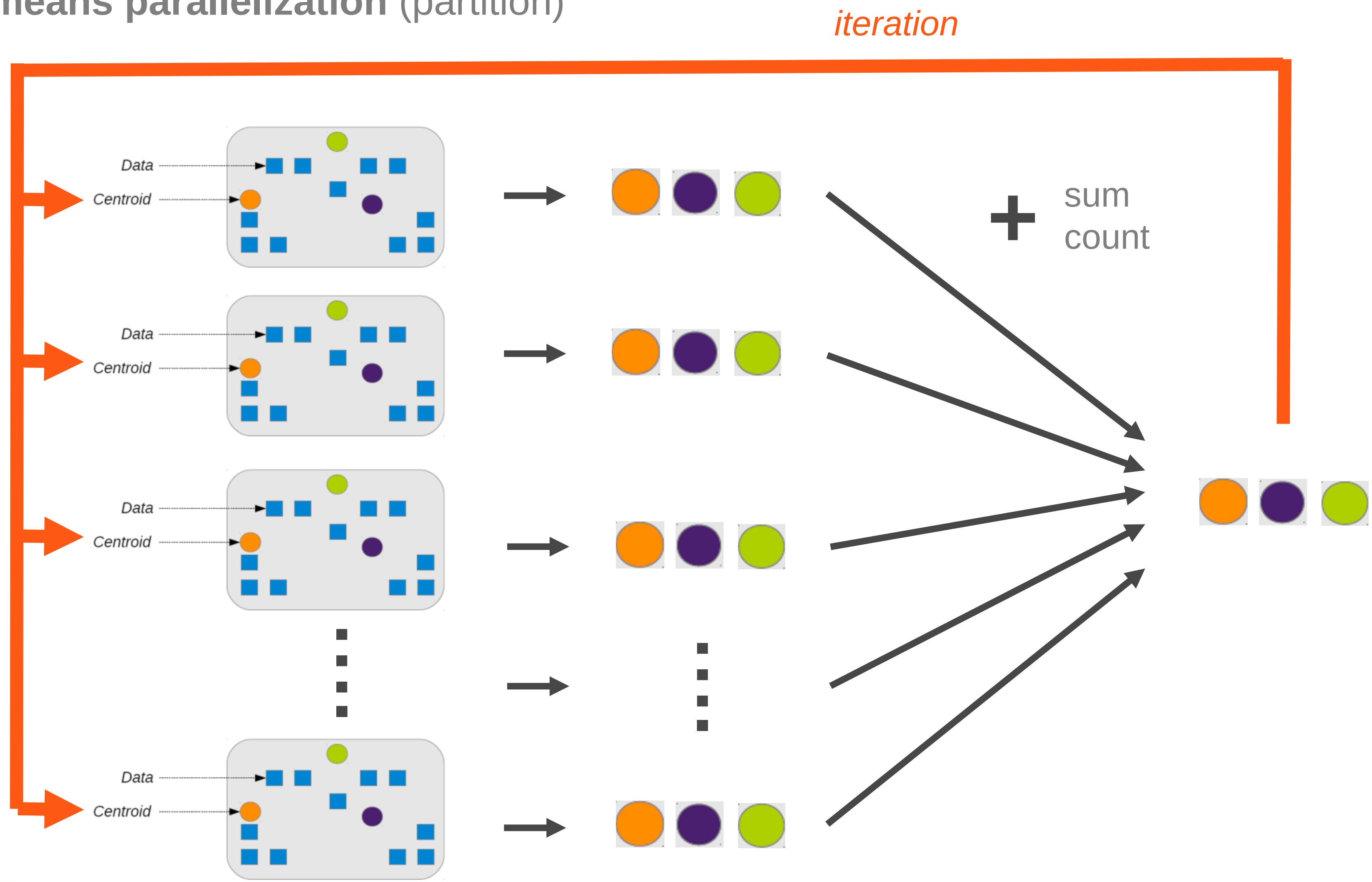
PRIVACY-PRESERVING DISTRIBUTED CLUSTERING

- **K-means algorithm** (centralized)



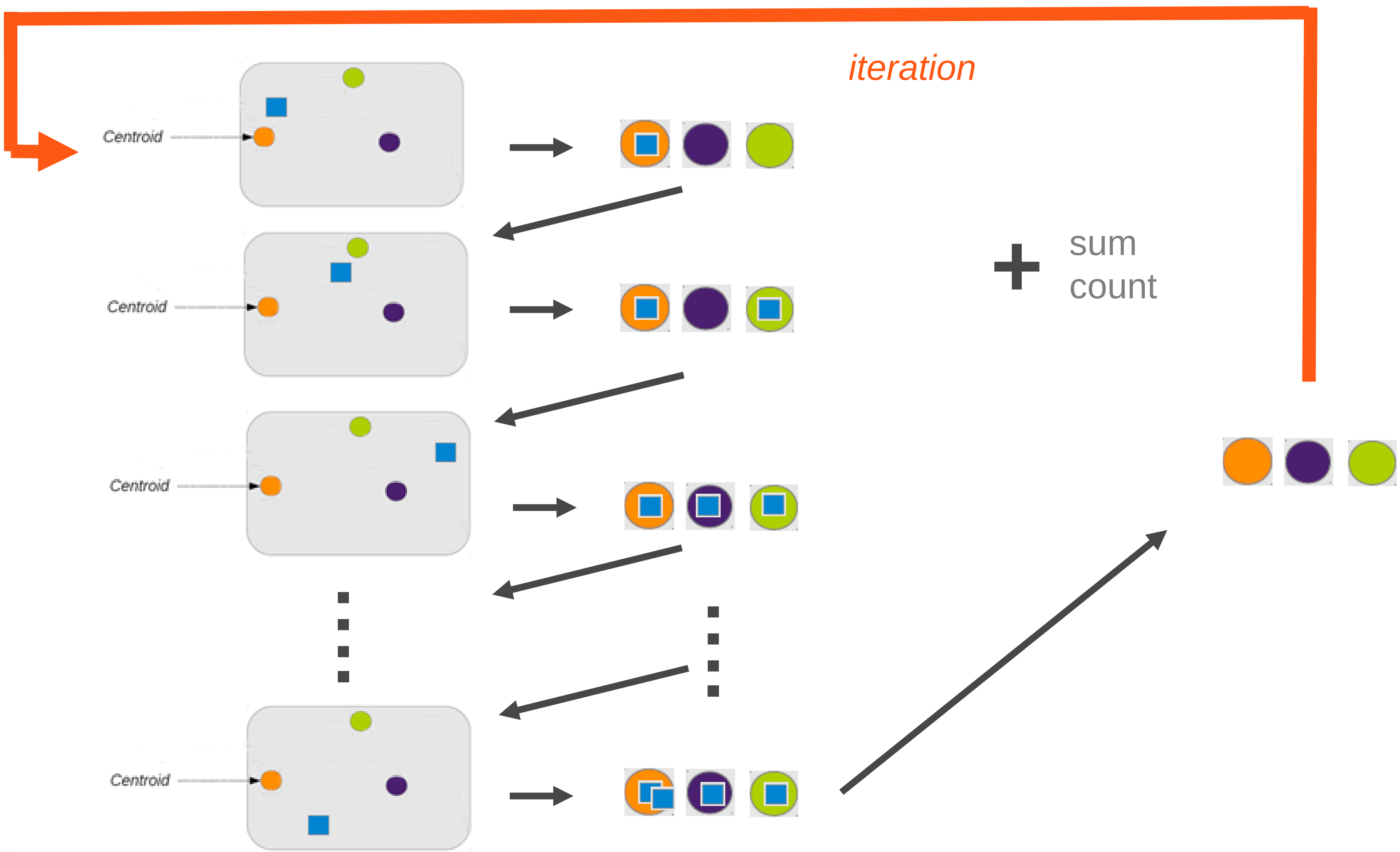
PRIVACY-PRESERVING DISTRIBUTED CLUSTERING

- **K-means parallelization** (partition)



PRIVACY-PRESERVING DISTRIBUTED CLUSTERING

- **K-means:** *circulation* of centroïds among individuals



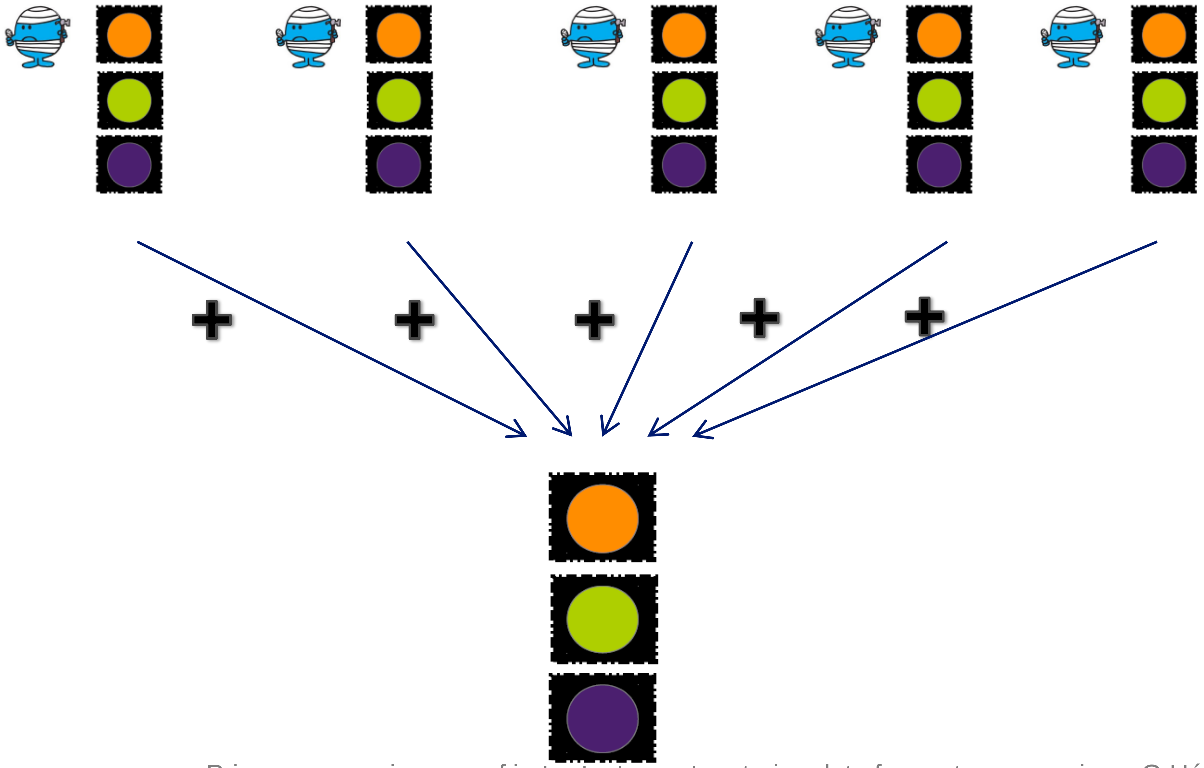
PRIVACY-PRESERVING DISTRIBUTED CLUSTERING

- **Circulation of 2 centroïd structures among individual participants**
  - Cleartext centroïds for local assignment of individual time series to the closest cluster
  - Encrypted centroïds built gradually from assignments for the next iteration

	
Cleartext centroids **perturbed** (differential privacy)	**Encrypted** means (additively-homomorphic)

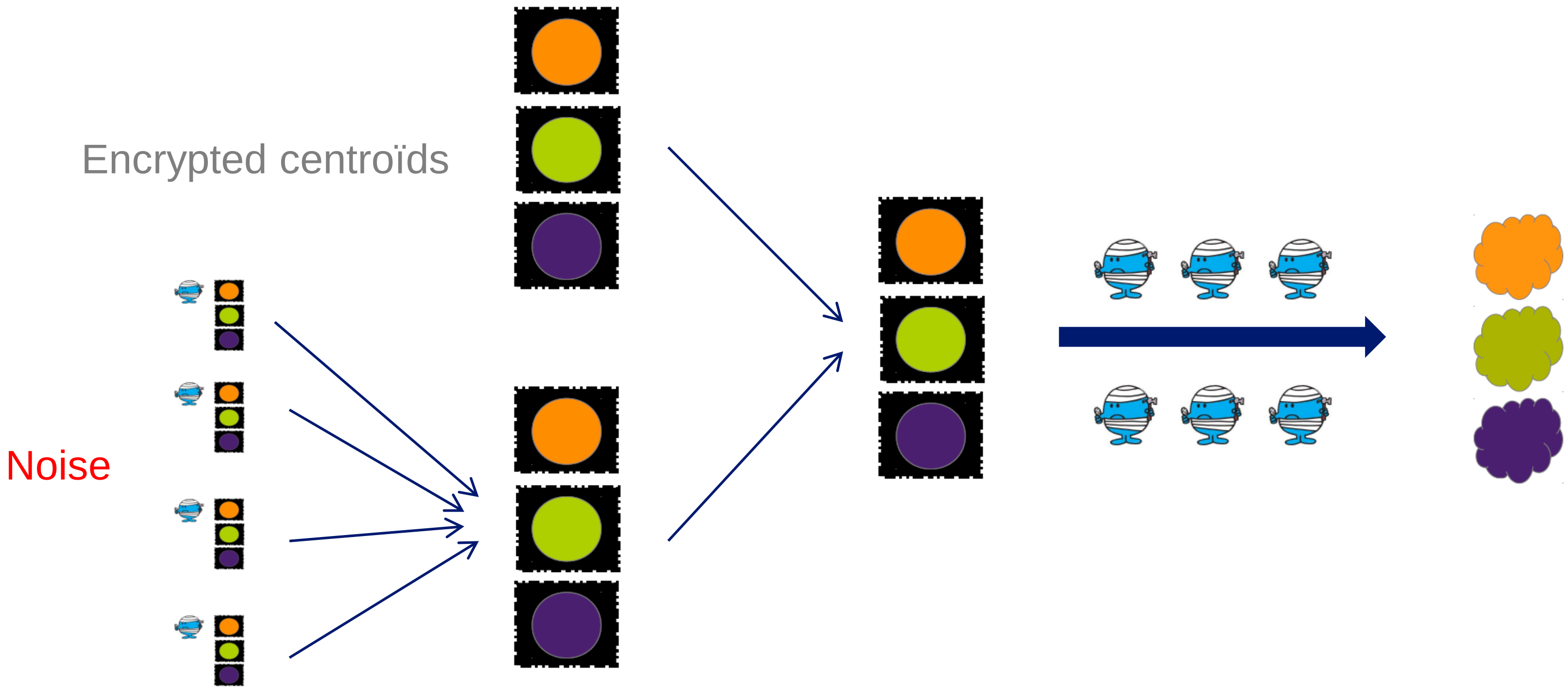
PRIVACY-PRESERVING DISTRIBUTED CLUSTERING

- **Centroïd computation within an iteration**
  - Two additive parts: SUM and COUNT
  - Use of additive homomorphic encryption *(allows addition directly on encrypted data)*



PRIVACY-PRESERVING DISTRIBUTED CLUSTERING

- **End of iteration**
  - Decryption of centroïds for the next iteration but:
    - Introduction of noise in centroïds before decryption (differential privacy)
  - Collaborative decryption



PRIVACY-PRESERVING DISTRIBUTED CLUSTERING

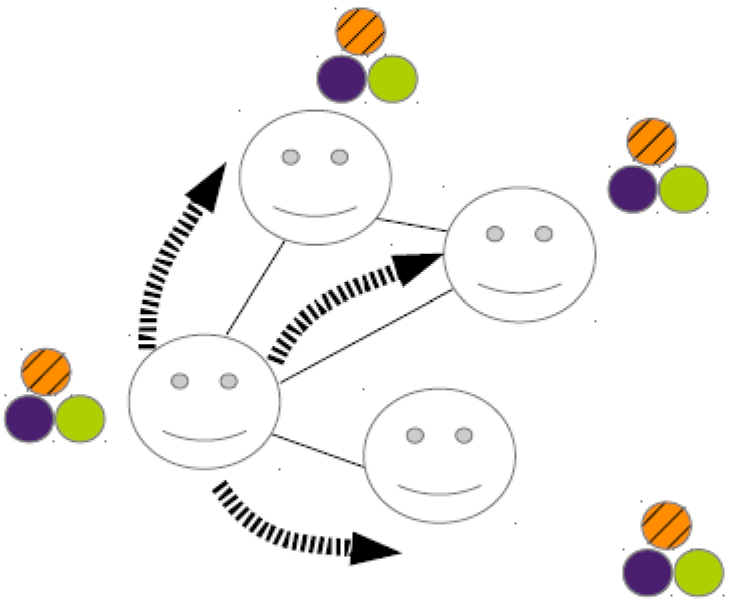
- **Association of house/equipment/occupants characteristics to clusters**
  - Last iteration
  - Counting for each combination *characteristic x cluster*
  - Similar protection: encryption + noise + collaborative decryption

PRIVACY-PRESERVING TIME-SERIES CLUSTERING

- **Privacy-preserving distributed clustering**
- **P2P infrastructure**
- **Evaluation**

PRIVACY-PRESERVING TIME-SERIES CLUSTERING

- **P2P (peer-to-peer) architecture**
  - No central server (local operations preserving privacy)
  - Scalability to millions of customers
  - Robustness to connections / disconnections (churn)
  - Sum computations using a « **gossiping** » algorithm
    - repeated averages between participants (adaptation of usual gossip sum algorithm)



PRIVACY-PRESERVING TIME-SERIES CLUSTERING

- **Privacy-preserving distributed clustering**
- **P2P infrastructure**
- **Evaluation**

PRIVACY-PRESERVING TIME-SERIES CLUSTERING

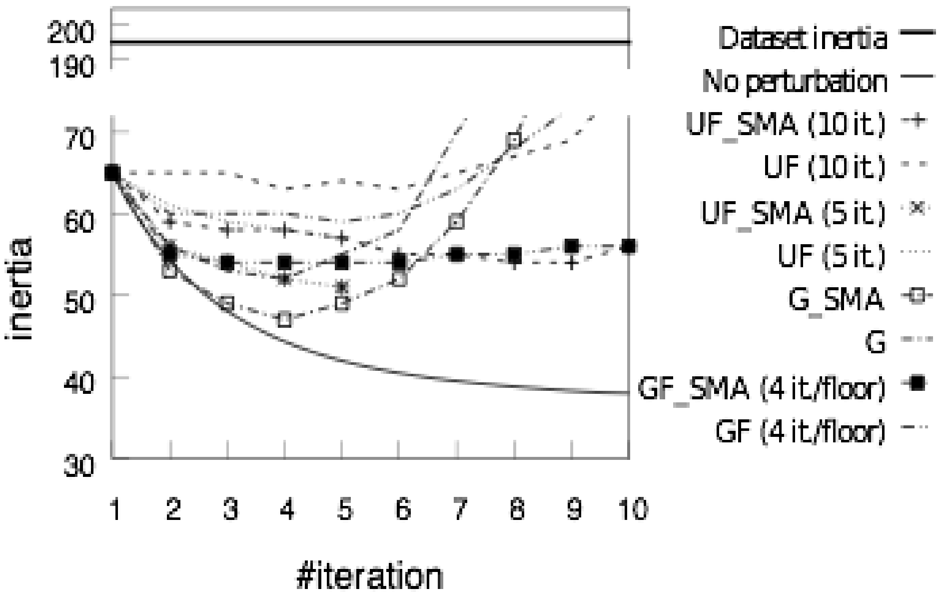
- **Evaluation questions:**
  - **Quality of clustering:**
    - Perturbed centralized k-means implementation
    - Measured by the intra-cluster inertia
    - Datasets : Irish CER (3M real electrical consumption time-series) and NUMED (1.2M synthetic tumor growth time-series)
  - **Latencies** of gossip algorithms: distributed computing simulator (Peersim)
  - **Local performances** (*i.e.,* CPU times, bandwidth consumption): laptop with *current average*+ resources

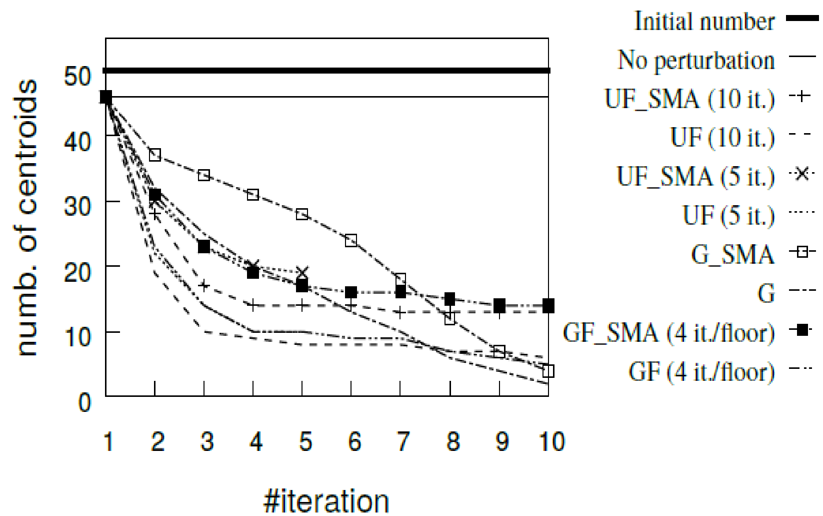
PRIVACY-PRESERVING TIME-SERIES CLUSTERING

- **Quality of clustering**
  - Varying participants for each iteration (connections/disconnections)
  - Introduction of noise
    - High perturbation for small clusters
    - Large clusters « eat » small clusters
  - Distribution of privacy budget between iterations
  - Smoothing time series after noise introduction
  - Early stopping

PRIVACY-PRESERVING TIME-SERIES CLUSTERING

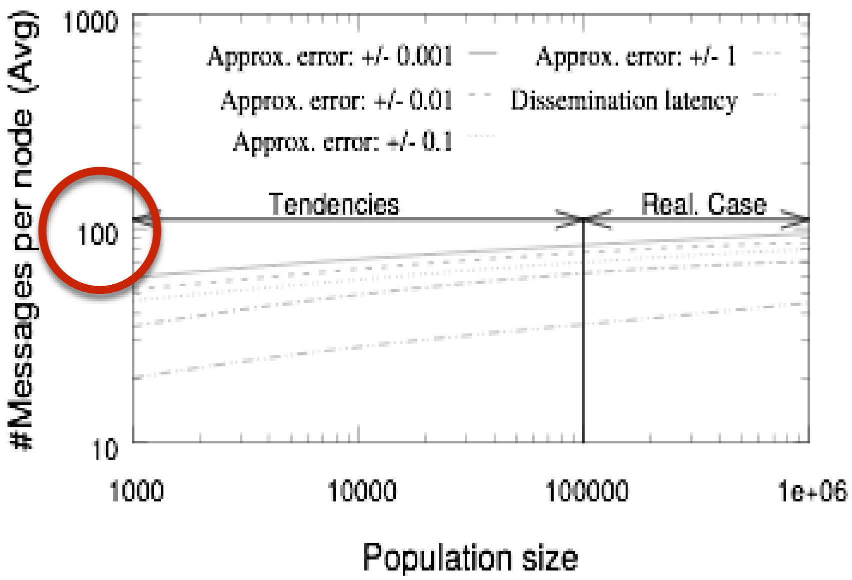
- **Quality of clustering: example of settings**
  - Clustering : k = 50 centroids, CER dataset, 24 numbers per time-series
  - Security : differential privacy budget ε = 0.69, encryption key length 1024 bits

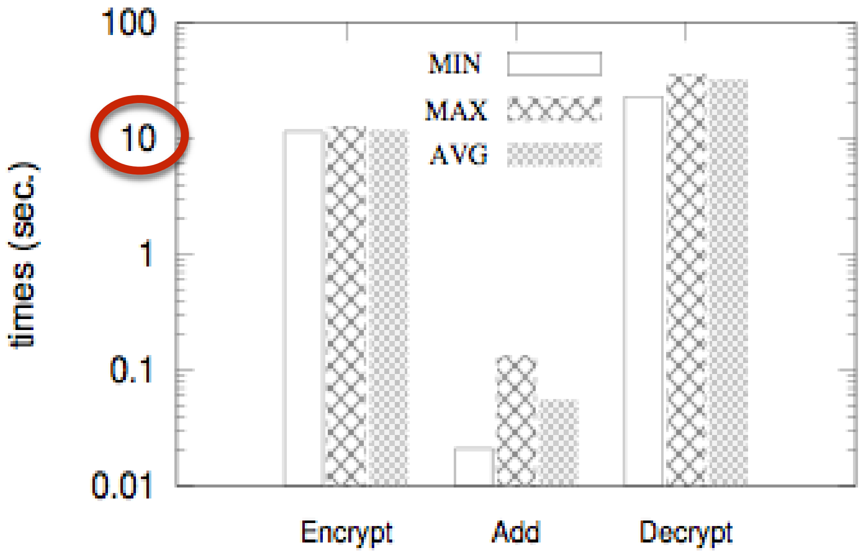




PRIVACY-PRESERVING TIME-SERIES CLUSTERING

- **Affordable communication and computation costs**





CONCLUSION

- **Chiaroscuro :**
  - First massively distributed privacy-preserving clustering solution for time series
  - Clustering: *k*-means-like algorithm (simplicity)
  - Distribution: Gossip-based (scalability and fault-tolerance)
  - Privacy: encryption and differential privacy

- **Future work :**
  - Functional representation of time series
  - Malicious participants
  - Other analytical algorithms

REFERENCES

“Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering”, T.Allard, G.Hébrail, F.Masseglia, E.Pacitti, Proceedings of the 2015 ACM SIGMOD.

"A New Privacy-Preserving Solution for Clustering Massively Distributed Personal Times-Series", T.Allard, G.Hébrail, F.Masseglia, E.Pacitti, Demonstration, ICDE, Helsinki, 2016.

“Differential privacy”, C. Dwork, in ICALP, 2006, p. 1–12.

“A Generalization, a Simplification and Some Applications of Paillier’s Probabilistic Public-Key System”, Damgaard et M. Jurik, in Proceedings of the 4th International Workshop on Practice and Theory in Public Key Cryptography: Public Key Cryptography, London, UK, UK, 2001, p. 119–136.

“Gossip-Based Computation of Aggregate Information”, D. Kempe, A. Dobra, et J. Gehrke, in FOCS, Washington, DC, USA, 2003, p. 482–491, 2003.