Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data

Tri Kurniawan Wijaya[1*], Tanuja Ganu[2], Dipanjan Chakraborty[2], Karl Aberer[1], Deva P. Seetharam[2]

[1)]EPFL, Switzerland & [2)]IBM Research, India

[*)]The work is done during the author's internship at IBM Research, India

supported by EU FP7 WATTALYST

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

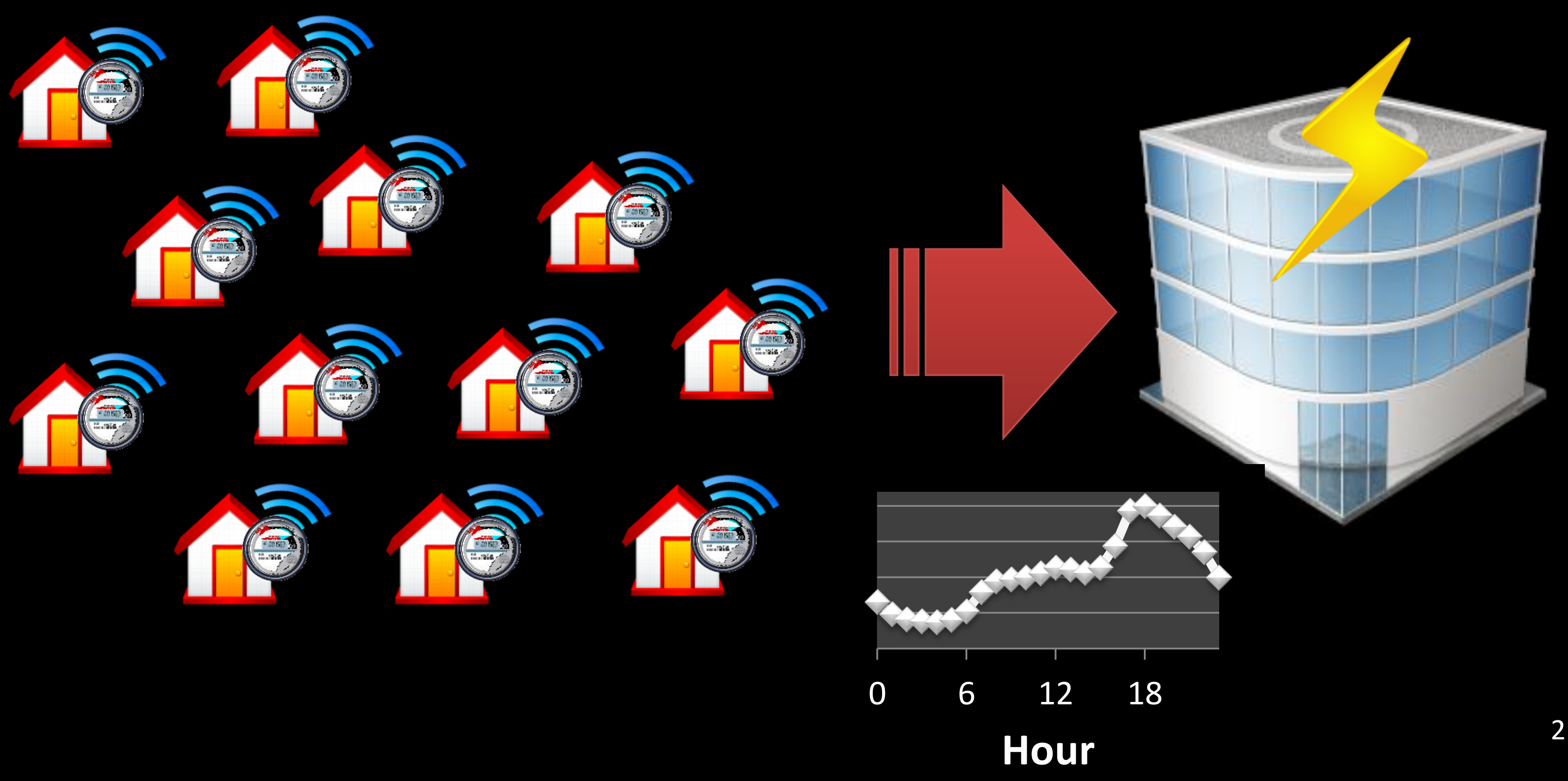
WATTALYST

IBM Research

Smart meters

measure energy consumption at homes

communicate the measurements to utility companies



Smart meters (2)

angels

demand response

match supply and demand

prevent black-out

renewable energy sources

theft detection

fault detection

demons

burglary

targeted marketing

privacy breach

insurance companies

press

Challenges

1 Versatile consumer segmentation framework

2 Determine behavioral change over time

3 Identify clusters' characteristics

1 Consumer segmentation

past

specific challenges

specific applications

adhoc

near future

general
versatile

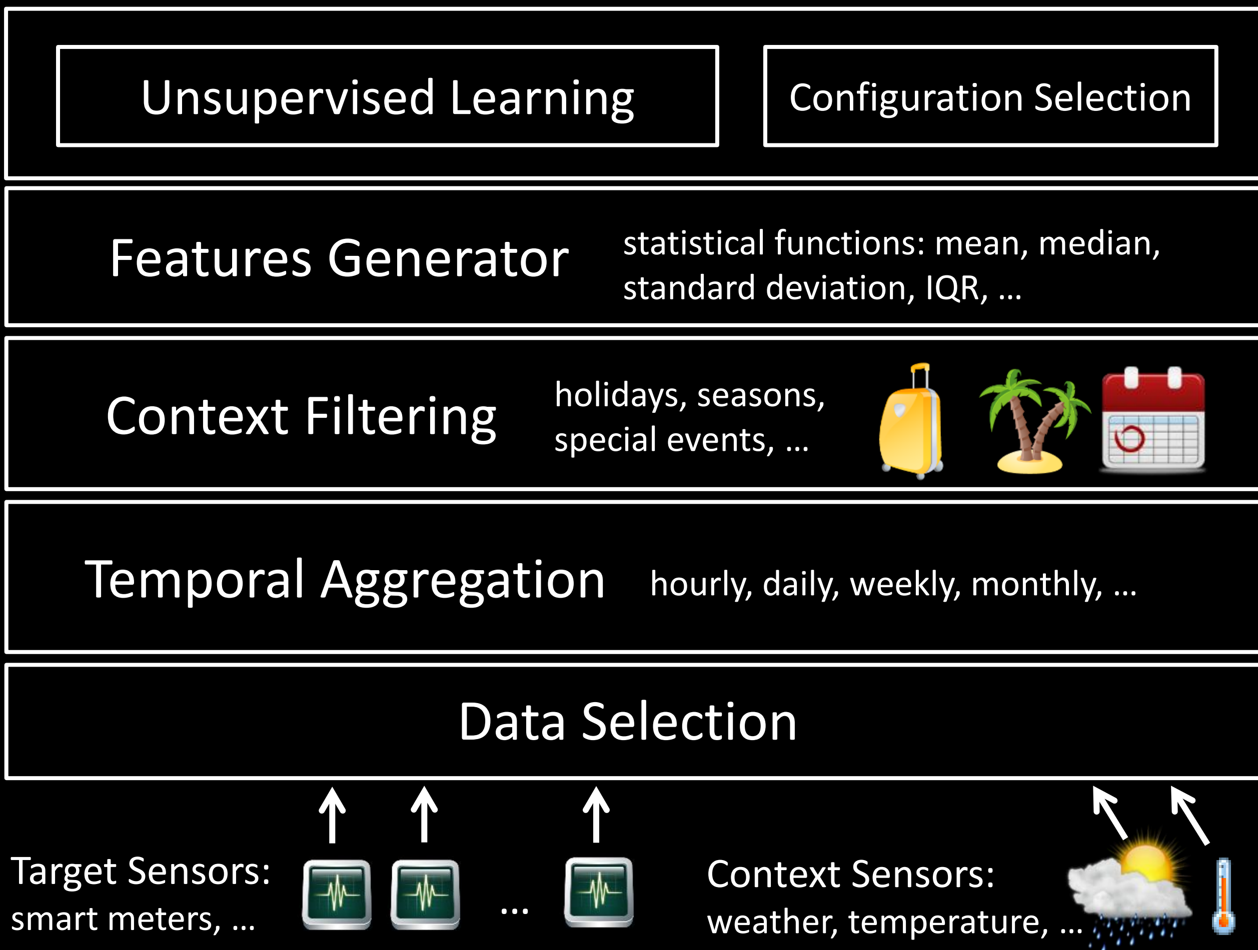
framework

quick analysis

context-aware

decision support

Our Framework



5 Customized algorithm

choose algorithm, #clusters (or auto)

4 Customized features

mean, std dev, IQR, median

3 Customized context

summer, winter, weekend, January, February, temp > τ

2 Customized temporal aggregation

hourly, every 3 hours, daily, weekly, or monthly

1 Customized data selection

period of time, subset of customers, time of day

See the complete formalization in our paper

2 Cluster consistency

Given all of these clusters, what do we want to know?

- Does this consumer change her cluster?
  note that: clusters are label-invariant

Individual to cluster consistency:

$$i2c(x, C_1, C_2) = \frac{|C_1(x) \cap C_2(x)| + \left|\left(X \setminus C_1(x)\right) \cap (X \setminus C_2(x))\right| - 1}{|X| - 1}$$

$$= \frac{\#(\text{friends} \ \rightarrow \ \text{friends}) + \#(\text{non}-\text{friends} \ \rightarrow \ \text{non}-\text{friends})}{|\text{customers}|}$$

consistent

inconsistent

the lower the value, the more likely x changes her cluster

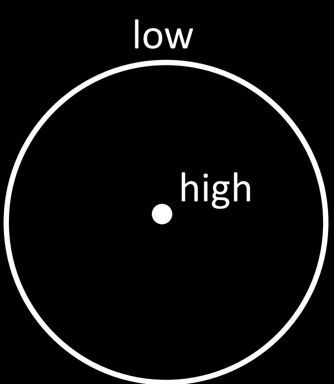
Clustering consistency index

- How far does this consumer change?

  distance rank

$$dr\big(x, C(x)\big) = \frac{\big|\{x' \mid dist\big(x, \zeta^{C(x)}\big) < dist\big(x', \zeta^{C(x)}\big), x' \epsilon C(x)\}\big|}{|C(x)|}$$

$=$ proportion of fellow cluster members who are farther from the centroid



the higher the value, the more confidence we are that x belong to C(x)

allows us to be (cluster) size-invariant

Clustering consistency index

- Does the cluster configuration changes?

  For example, over time?

  Cluster to cluster consistency:

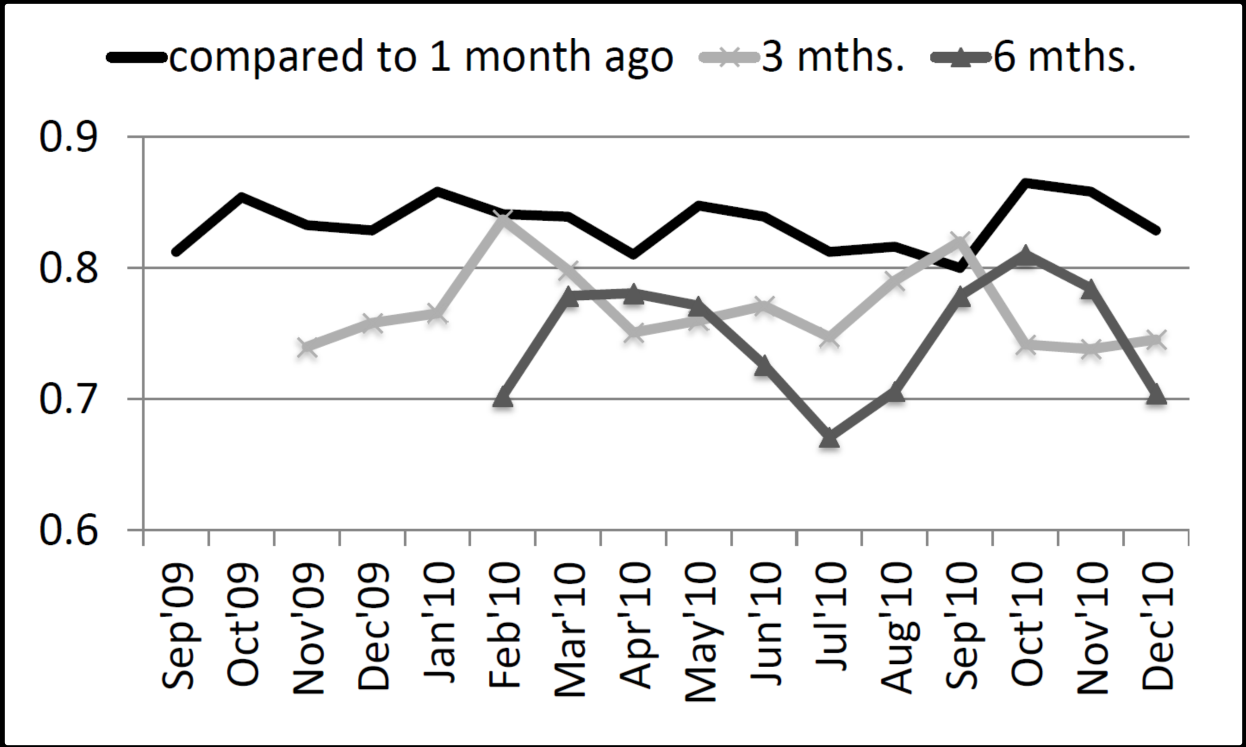
$$ccc(C_1, C_2, X) = \frac{1}{X} \sum_{x \,\epsilon X} i2c(x, C_1, C_2)$$

the lower the value, the higher the difference between C_1 and C_2

consistent ↑

inconsistent

Cluster to cluster consistency

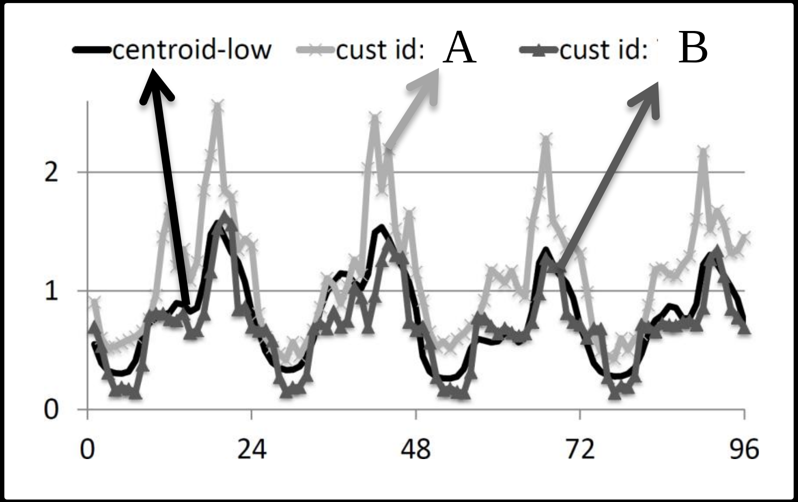


the higher, the more similar

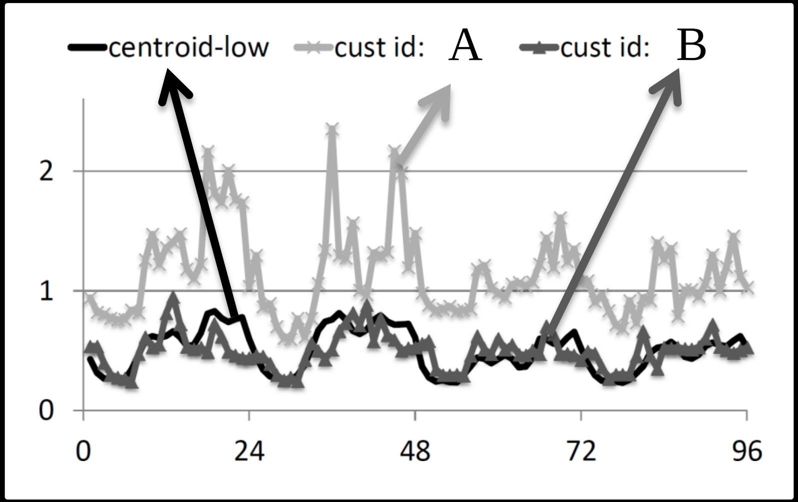
See comparison with 6 months ago:

- There are not so much difference between autumn and spring.
- But, there are a lot of difference between summer and winter.
- Next slide, more on Jan vs Jul …

Individual to cluster consistency



Jan



Jul

- In Jan, **A** and **B** are in the low consumption cluster
- *i2c(* **A**, *Jan, Jul)* = low (changes her cluster) → *dr(A, Jul)* = high
- *i2c(* **B**, *Jan, Jul)* = high (stays in the low consumption cluster)
- Devise a personalized energy (saving) feedback for **A**! While her "friends" reduce their consumption in Jul (summer), **A** did not!

3 Knowledge extraction

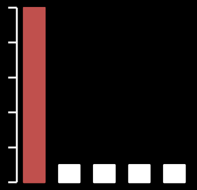
- What are the characteristics that define a cluster?

  get insight from the survey data (consumer characteristics)

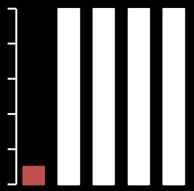
  How discriminative is *(q,a)* to cluster *c*?

$$DI_c(q,a) = \frac{\#_c(q,a) \; - \; \#_{\neg c}(q,a)}{\max\{\#_c(q,a), \#_{\neg c}(q,a)\}}$$

DI > 0, discriminative *positive*



DI < 0, discriminative *negative*



Clusters’ characteristics

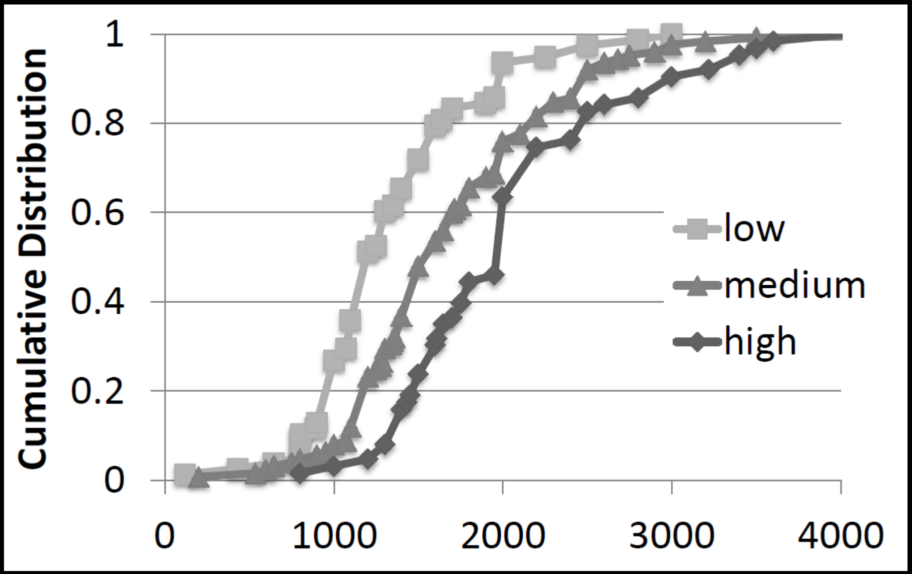
Clusters based on absolute consumption

Cluster	Question	Answer	DI
low	family type	single	0.86
	floor area (sq ft.)	805-1073	0.86
	#bedrooms	≤ 2	0.85
medium	electric shower	(-) ≥ 20 mins	-0.76
	family type	(-) single	-0.61
	floor area (sq ft.)	2300-2750	0.56
high	#children	≥ 4	0.93
	family type	(-) single	-0.90
	floor area (sq ft.)	(-) 1200	-0.87

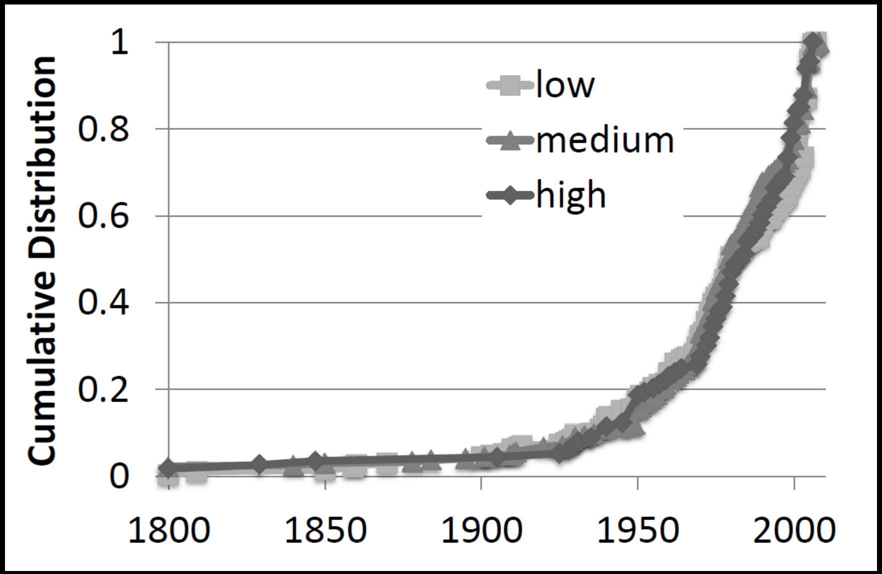
Clusters based on consumption variability

Cluster	Question	Answer	DI
low	water pump	(-) 1-2hrs	-0.88
	family type	single	0.80
	washing machine	(-) 2-3 loads	-0.76
medium	electric shower	10-20 mins	0.59
	family type	(-) single	-0.55
	#children	(-) ≥ 3	-0.54
high	tumble dryer	≥ 2 to 3 loads	0.90
	#children	≥ 4	0.88
	floor area (sq ft.)	2800	0.79

Floor area vs year built



Floor area



The year the houses were built

CDFs are clearly distinguishable

CDFs are coincides to each other

Appliance ownership for DI ≥ 0.60

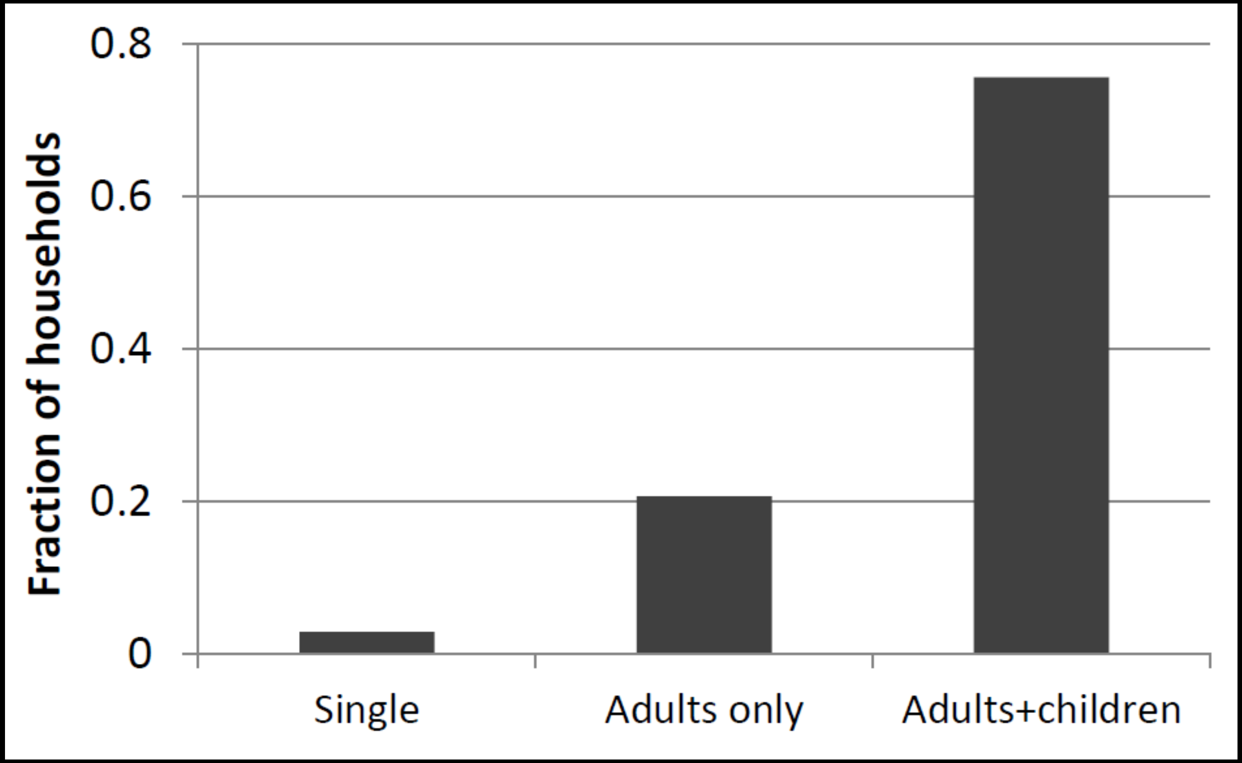
Clusters based on <u>absolute consumption</u>

#	Appliance	Cluster	Ownership	DI
1	dishwasher	high	(-) no	-0.76
2	games consoles	low	(-) yes	-0.70
3	tumble dryer	low	no	0.68
4	dishwasher	low	no	0.67
5	games consoles	high	yes	0.61

Clusters based on <u>consumption variability</u>

#	Appliance	Cluster	Ownership	DI
1	dishwasher	high	(-) no	-0.72
2	tumble dryer	high	(-) no	-0.72
3	tumble dryer	low	no	0.71
4	games consoles	low	(-) yes	-0.69
5	dishwasher	low	no	0.67
6	games consoles	high	yes	0.60

Games consoles



Fraction of households which own games consoles

Since family type is highly discriminative for consumer energy consumption behavior, this correlation might explain why games consoles ownership is also highly discriminative.

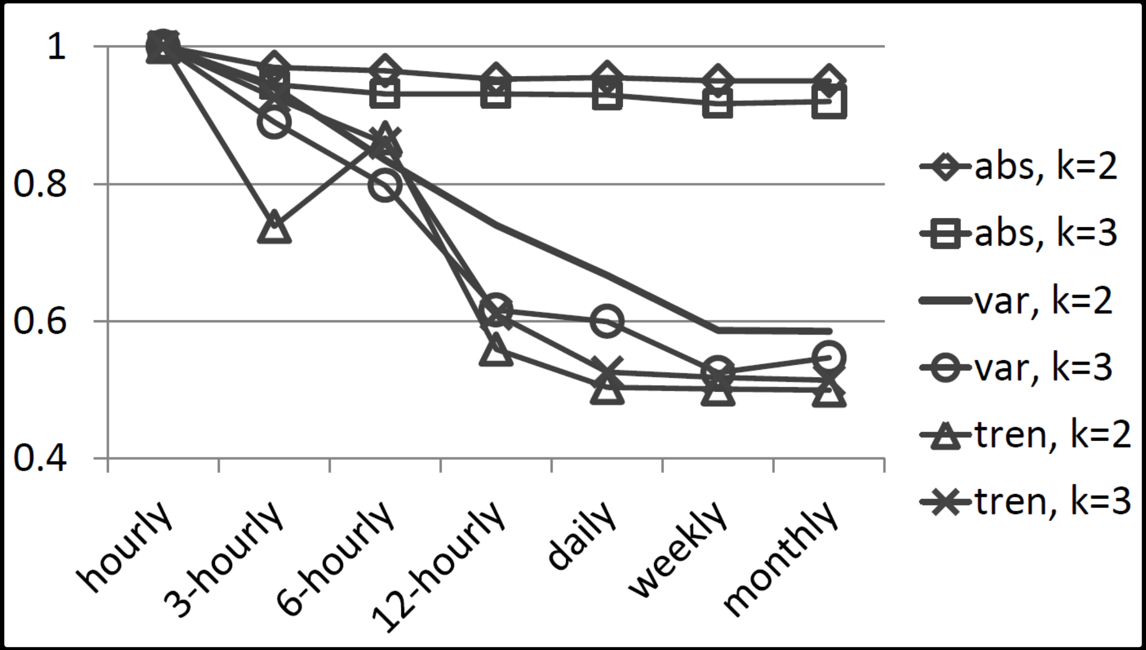
Conclusion

- Versatile, context-aware consumer segmentation framework
  - temporal aggregation, context filtering, feature generation
- Cluster consistency index
  - Which consumers change their clusters? How far?
  - track clusters' changes over time
- Discriminative index
  - Clusters : unsupervised learning;
    - It is imperative to understand what they are made of, extract the main characteristics which define the clusters.

end of presentation

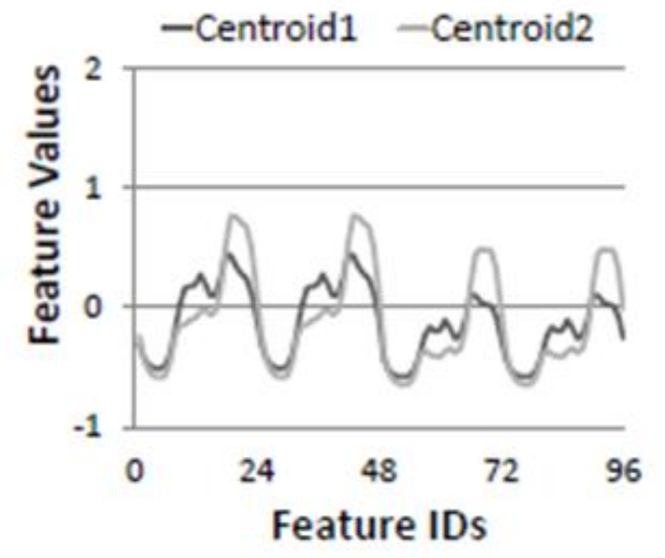
Cluster to cluster consistency

can be used to find out the effect of temporal aggregations on the consumer segmentation results



the higher, the more similar

Consumer segmentation based on …	Do temporal aggregation granularities matter?
Absolute consumption	✗
Consumption variability	✓
Consumption trends (or "shape")	✓


Centroid1
Centroid2
Feature Values
2
1
0
-1
0
24
48
72
96
Feature IDs
(c) Trends-all

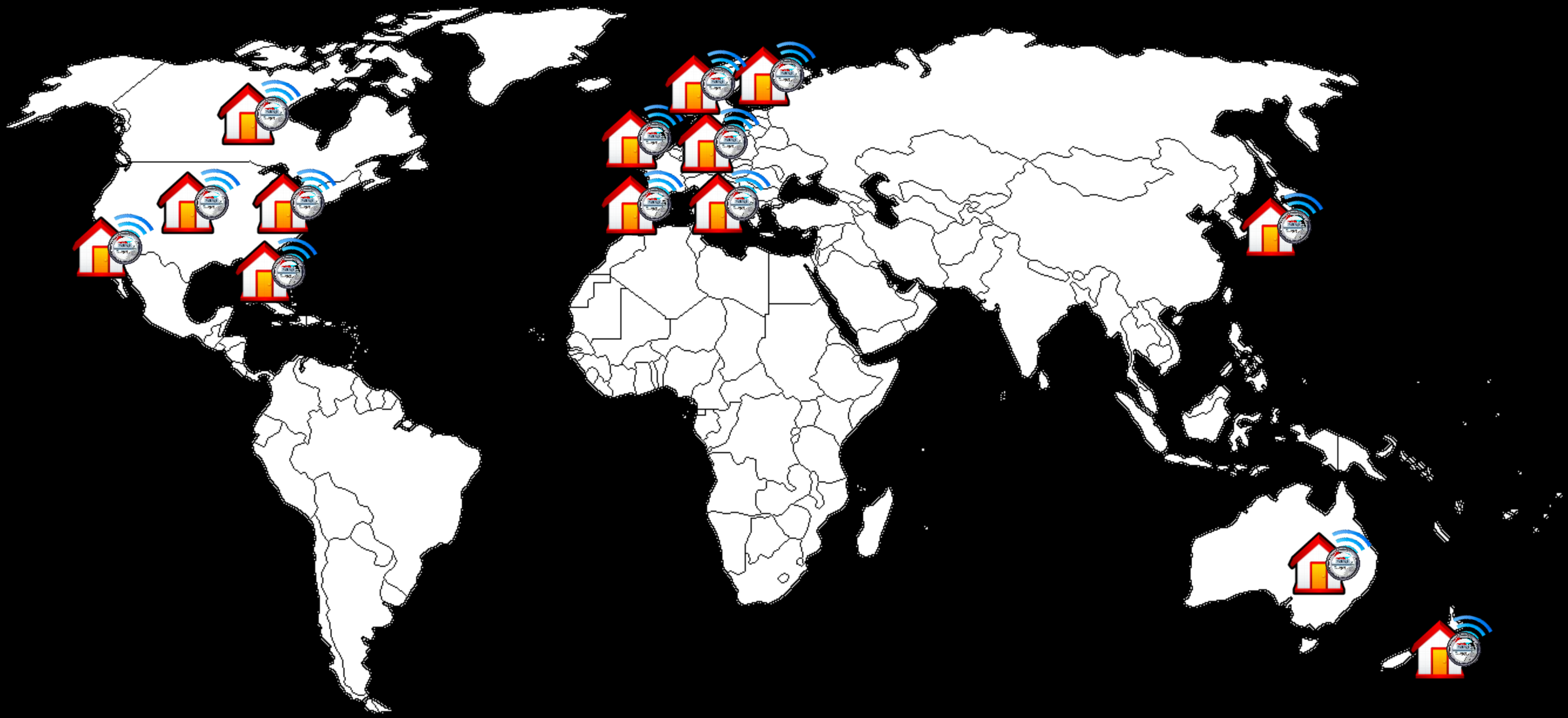
Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data

Tri Kurniawan Wijaya[1], Tanuja Ganu[2], Dipanjan Chakraborty[2],
Karl Aberer[1], Deva P. Seetharam[2]
[1]EPFL, Switzerland & [2]IBM Research, India

supported by EU FP7 WATTALYST

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Worldwide deployment



Automatic cluster configuration selection

- From a set of cluster configuration
  - Rank all configurations using the
    - Silhouette,
    - Dunn, and
    - Davies-Bouldin indices
  - Majority voting using the three (ranked) lists
    - using the 1st rank from each list
    - if the majority is not found, continue to the 2nd (3rd , 4th …) until the majority is found or the lists are exhausted

Jan

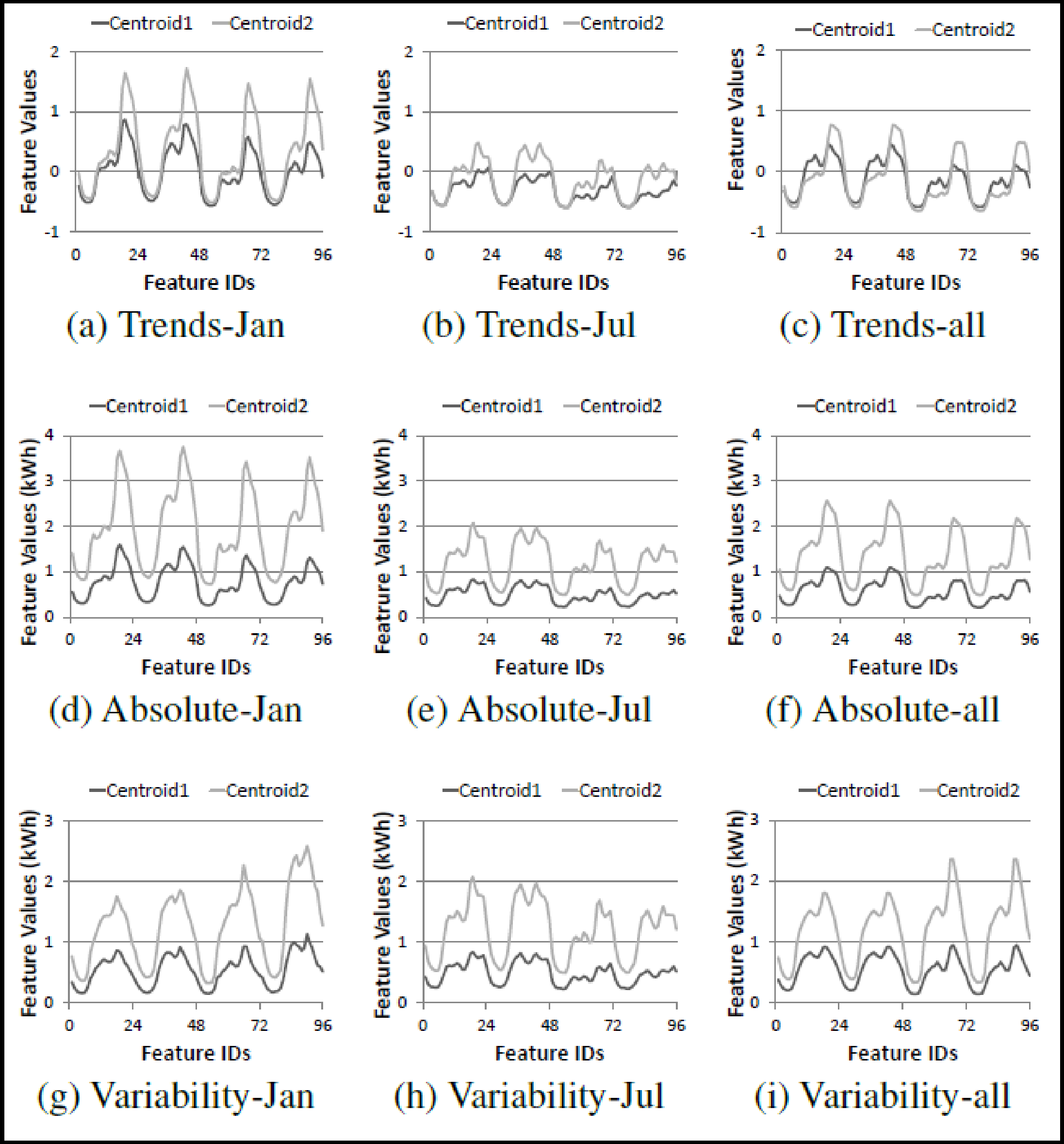
Jul

All year long

trends

absolute

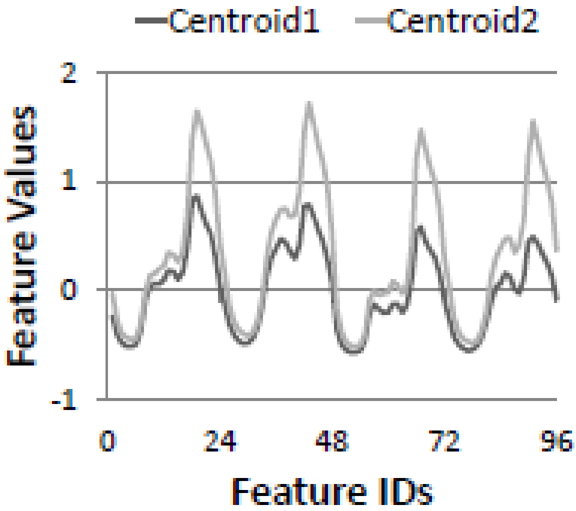
variability



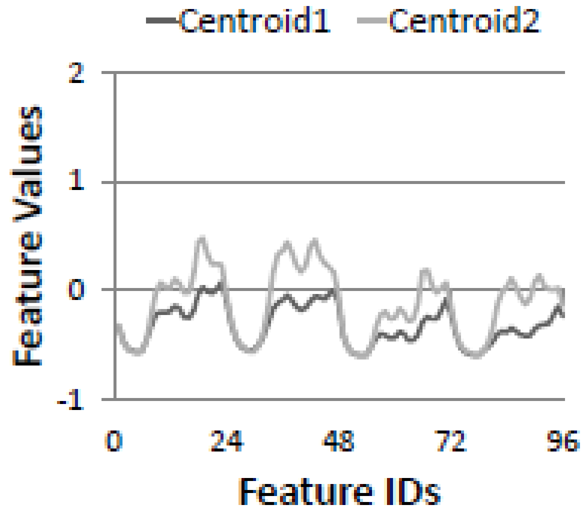
(a) Trends-Jan (b) Trends-Jul (c) Trends-all

(d) Absolute-Jan (e) Absolute-Jul (f) Absolute-all

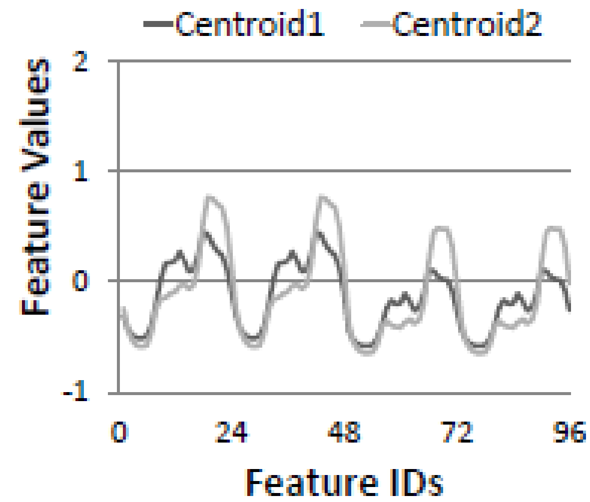
(g) Variability-Jan (h) Variability-Jul (i) Variability-all



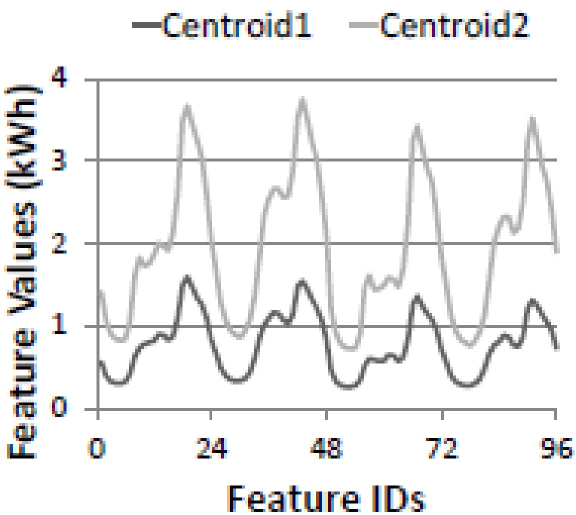
(a) Trends-Jan



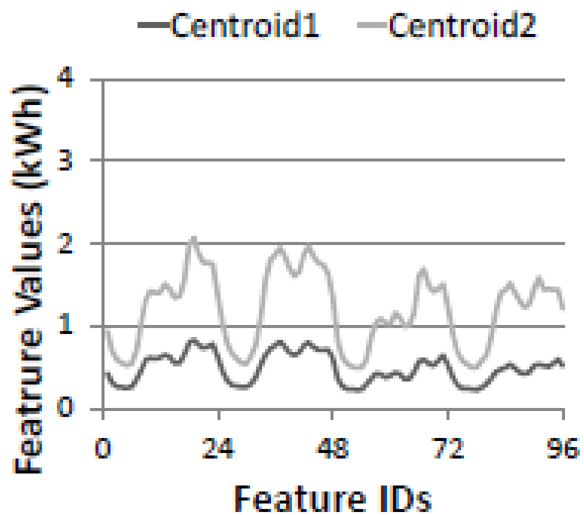
(b) Trends-Jul



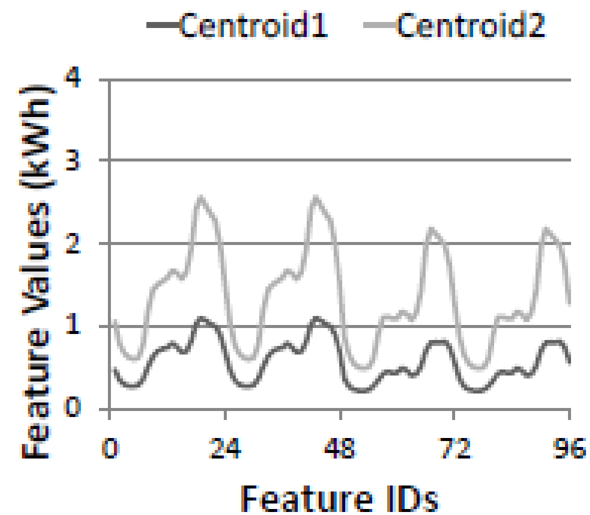
(c) Trends-all



(d) Absolute-Jan



(e) Absolute-Jul



(f) Absolute-all

Numerical/ordinal questions

- how many … ?
- approximate floor area ?

- special treatment
- introducing splitting points:
  – how many?
  – where to put?

  combinatorial problem
- solution:
  – sort answers ascending
  – create ranges from n-gram of answers