Sample-based Clustering for Big Data using Coresets

Le Hong Trang

Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology, VNU-HCM
lhtrang@hcmut.edu.vn

Lab of Software Architecture and Information System (lasaris)
Faculty of Informatics, Masaryk University
March 2019, Brno, Czech Republic

Outline

HCMC University of Technology

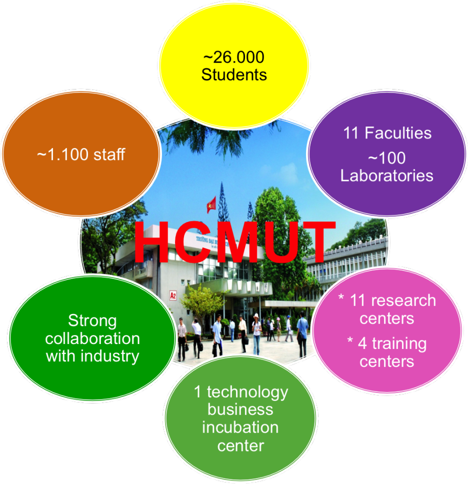
Location

HCMC University of Technology

Campus

HCMC University of Technology

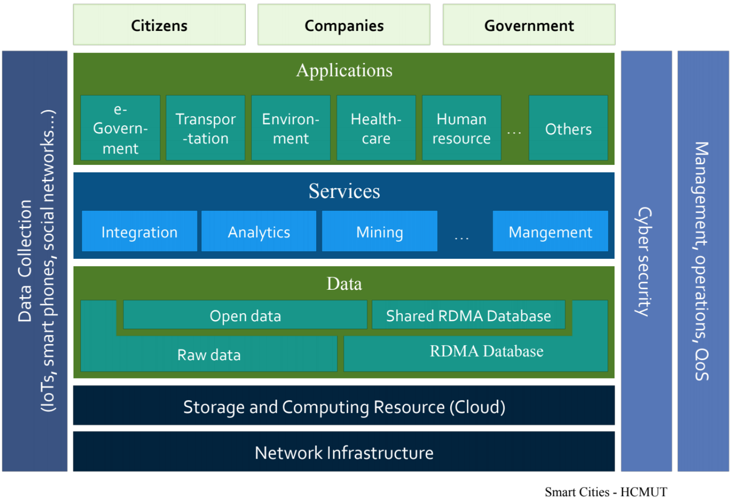
Key facts



Outline

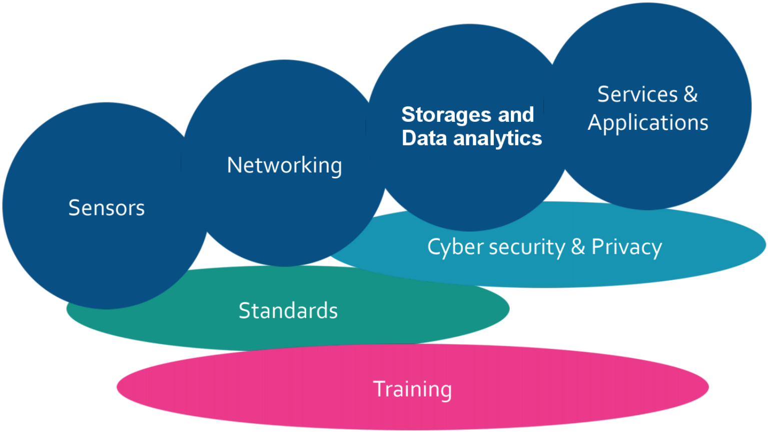
An ICT Architecture for Smart Cities

Overview



An ICT Architecture for Smart Cities

Research topics

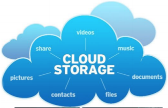


An ICT Architecture for Smart Cities

Data analytics

Data analytics

Learn



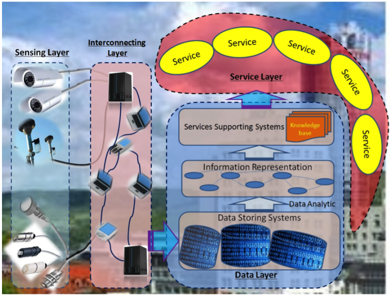
Collect

Data analytics

Big data

- A smart city will developed on an IoT infrastructure.
  - It should be network of sensors, devices, and citizens.
- A mount of data will be generated
  - huge size,
  - complicated structure,
  - continuously and fastly generated,
  - and so on.

  Called Big Data.

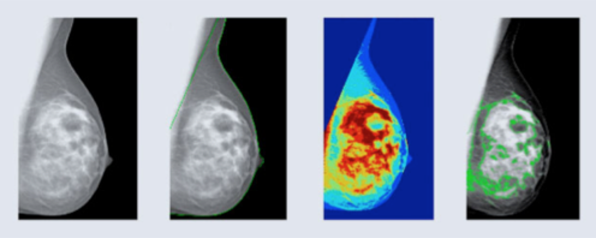


[Sun et al., 2015]

Big Data Clustering

where?

- Economy,
- biology,
- Medicine,
- Transportation,
- Education.



[Guillaume Agis's blog]

The role of big data clustering

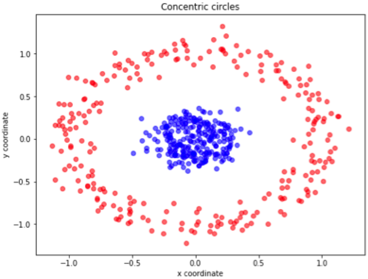
- **In order to understand and explore the structure of the data for analysis purpose.**

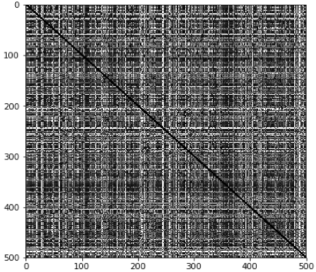
Challenges in Big Data

(a) Huge size (volume)

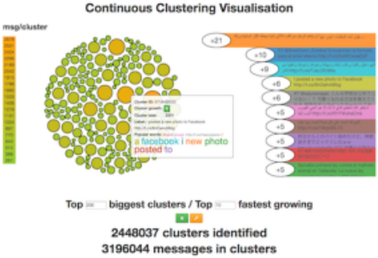
- Large number of data object: computational cost increased exponentially.
- High dimension: curse of dimensionality.

(b) Many types of data (variety).





Challenges in Big Data



[CeADAR, Dublin]

(c) Continuously generated (velocity)

- Real time processing.
- Deal with streaming data.

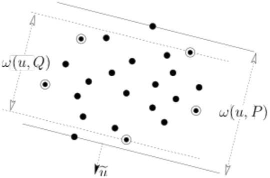
Sampling Method

Coreset concept [Agarwal et al., 2004]

- Proposed for geometric approximation of a set of points in $\mathbb{R}^d$.
  - Given a set T and $\varepsilon > 0$, let μ be a *monotonic function* defined on T, that is, for $S \subseteq T$, $\mu(S) \leq \mu(T)$.
  - *Then, S is an ε-coreset of T w.r.t μ, if*

$$(1-\varepsilon)\mu(T) \leq \mu(S).$$

- $\omega(u, P) = \max_{p \in P}\langle u, p\rangle - \min_{p \in P}\langle u, p\rangle$ is an example for μ, where u is an arbitrary direction of P.



Sampling Method

Coreset for clustering [Har-Peled et al., 2004]

Definition

A set S of s points is an (k, ε)-coreset for a set T of $n > s$ points if

$$(1-\varepsilon)Cost_T(C) \leq Cost_S(C) \leq (1+\varepsilon)Cost_T(C),$$

for $C = \{c_1, c_2, \ldots, c_k\}$ a set of k centers.

- For a clustering problem, functions $Cost$ can be defined by

$$Cost_T(C) = \sum_{i=1}^{n} d(x_i, c_i^*) \text{ and } Cost_S(C) = \sum_{i=1}^{s} w_j d(y_i, c_i^{*\prime}).$$

where, $c_i^*, c_i^{*\prime} \in C$ respectively are closest centers for $x_i \in T$ and $y_j \in S$, i.e., $d(x_i, c_i^*)$ and $d(y_i, c_i^{*\prime})$ are minimum among k centers, $w_j = |T(y_j)|$, i.e., the number of items of T whose closest point in S is y_j.

Sampling Method

ProTraS [Ros and Guillaume, 2018]

1. Add new sample in the group with highest probability of cost reduction that combines
   - density-based probability: $P_{dens}(j) = \frac{w_j}{\max_i w_i}$,
   - distance-based probability: $P_{dist}(j) = \frac{d_j}{\max_i d_i}$.
2. Assign each pattern to the nearest sample.
3. Compute $Cost$.
4. If $(Cost > \varepsilon)$ goto Step 1.

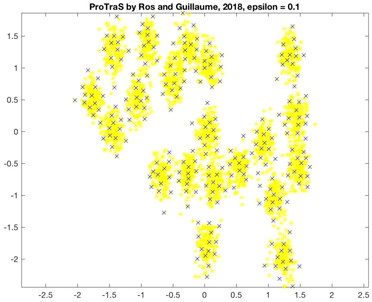
Theorem

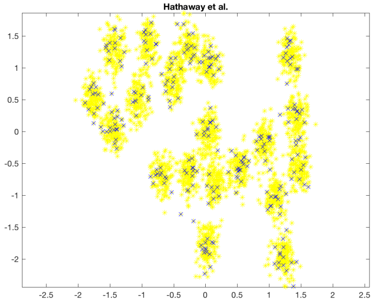
ProTraS yields a (k, ε)-coreset with

$$\varepsilon = \frac{\sum_{j=1}^{s} w_j d_j}{Cost_T(C)}.$$

Sampling Method

ProTraS vs. siVAT

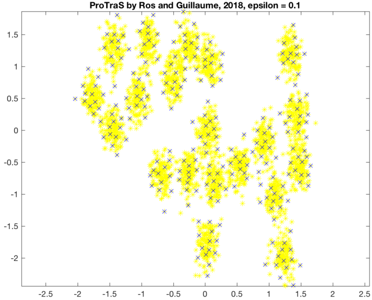


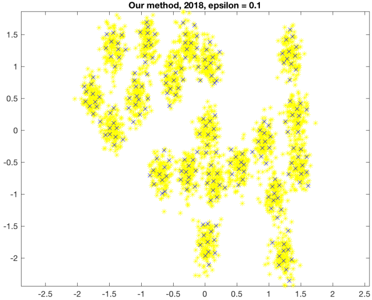


- Sample obtained by ProTraS is higher representative, compared with that by siVAT.
- But
  - uniformly distributed → difficult to highlight clusters in the sample.
  - may include noises and outliers.

Sampling Method

ProTraS: our improving



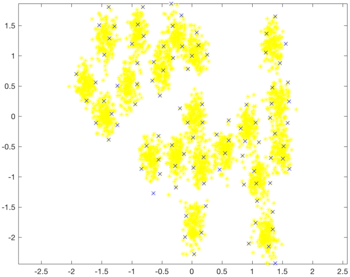


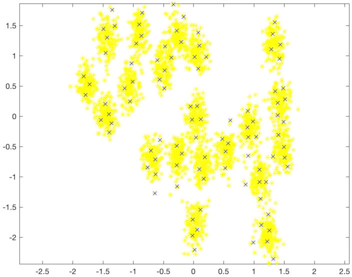
- Replace every representative point in the sample by the center of group represented by it.
  - Objects located at the boundary side of clusters will be replaced by interior ones of those.
  - New obtained sample thus should has separated clusters.

→ *obtain higher accuracy in VAT problem.*

Experiments

Comparison between ProTraS and our sampling





ProTraS vs. our sampling

Experiments

Sample sizes with different values of ε

Table 1: Sample size with $\epsilon = 0.1$ and 0.2.

Ord.	Dataset	Data size (T)	Sample size (S)		Ratio S/T (%)	
			$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.1$	$\epsilon = 0.2$
1	A.set 1	3000	261	97	8.7	3.23
2	A.set 2	5250	315	116	6	2.21
3	A.set 3	7500	341	119	4.55	1.59
4	FLAME	240	166	90	69.17	37.5
5	Birch-set 3	100000	424	153	0.424	0.153
6	JAIN	373	108	56	28.95	15.01
7	S.sets 1	5000	237	96	4.74	1.92
8	S.sets 2	5000	327	120	6.54	2.4
9	S.sets 3	5000	422	155	8.44	3.1
10	S.sets 4	5000	448	166	8.96	3.32
11	Dim sets 1	1351	17	10	1.26	0.74
12	Dim sets 2	2701	17	11	0.63	0.41
13	Dim sets 3	4051	20	8	0.49	0.2
14	Dim sets 4	5401	416	17	7.7	0.31
15	Dim sets 5	6751	379	19	5.61	0.28
16	data5k-CS	5000	44	17	0.88	0.34
17	data5k-NonCS	5000	264	95	5.28	1.9
18	data10k-CS	10000	25	10	0.25	0.1
19	data10k-NonCS	10000	114	40	1.14	0.4
20	data15k-CS	15000	61	22	0.41	0.145
21	data15k-NonCS	15000	111	44	0.74	0.293
22	data100k-10	100000	103	45	0.103	0.045
23	data100k-25	100000	191	73	0.191	0.073
24	data100k-27	100000	187	79	0.187	0.079
25	data200k-5	200000	108	44	0.054	0.022
26	data200k-17	200000	162	62	0.081	0.031
27	data1M	1000000	315	107	0.0315	0.0107
28	data1M-7	1000000	84	41	0.0084	0.0041
29	data1M-15	1000000	142	60	0.0142	0.006
30	data1M-55	1000000	355	131	0.0355	0.0131
31	data2M-77	2000000	457	159	0.023	0.008

Outline

Visual Assessment of Cluster Tendency

Clustering

- Notes
  - Most of proposed techniques concentrate on how to separate objects into proper groups.
  - Many algorithms, for example the family of k-means, require the number of clusters as an input.
  - Knowing an approximate number of clusters can help a clustering algorithm not only to speed up the process, but also to enhance its accuracy.
- *It is important to estimate a number of clusters before applying a suitable technique for the cluster analysis.*

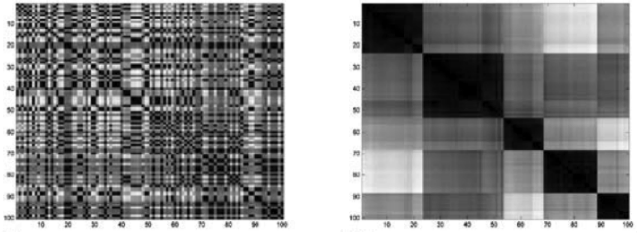
Visual Assessment of Cluster Tendency

(VAT)

- VAT: introduced by Bezdek and Hathaway, 2002.
  - Determine whether cluster are presents in a given dataset.
  - Visualize cluster structures in relational matrices among objects of the dataset.
- Main idea
  - Rearranges unlabled objects so that similar ones will be located nearby.
  - Highlights the cluster structure of a dataset in an intuitive image.

Visual Assessment of Cluster Tendency

VAT: main idea



$I(D)$ vs. $I(D^*)$

- Take a pairwise dissimilarity matrix of a dataset D ($I(D)$).
- Determine a potential partition of the dataset by Prim's algorithm.
- Reorder matrix D into D^* due to the obtained partition.
- Visualize D^* by a grayscale image $I(D^*)$.
- The *cluster tendency* is indicated by the "dark blocks" along the diagonal.

Visual Assessment of Cluster Tendency

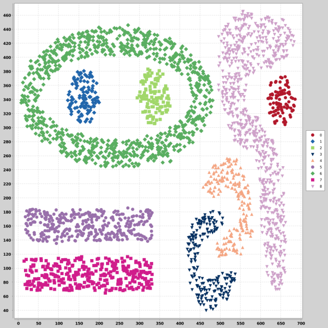
The VAT algorithm: variants

- Some variants were proposed to deal with datasets of irregular structure and large size. Some typical ones of them include
  - sVAT [Hathaway et al., 2006]: scalable VAT for large datasets *using sampling*.
  - iVAT [Wang et al., 2010]: improved VAT for datasets of complicated structure *using a path-based distance*.
  - Revised iVAT [Havens and Bezdek, 2012]: improve the computation of the path-based distance in iVAT.
  - Combining sVAT and iVAT to obtain *siVAT*.

Visual Assessment of Cluster Tendency

The VAT algorithms: difficulties

- Sampling for large datasets
  - Need an overestimate of the true but unknown number of clusters.
  - Sample points are chosen randomly.
  - → Low representativeness.



A complex dataset with 9 clusters.

Sample-based VAT Method

Proposed algorithm

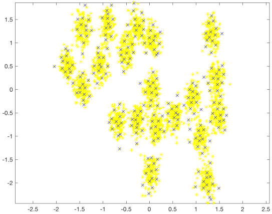
Input: $T = \{x_i\}$, for $i = 1, 2, \ldots, n$, a tolerance $\varepsilon > 0$.
Output: A sample S and D'^*.

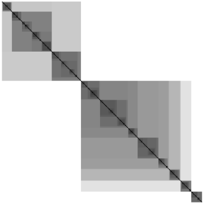
1: Call ProTraS for T and ε to obtain $S = \{y_j\}$ and $P(y_j)$.
2: $S' = \emptyset$.
3: **for all** $y_j \in S$ **do**
4: $\quad y_k^* = \text{argmin}_{y_k \in P(y_j)} \sum_{y_l \in P(y_j)} d(y_k, y_l)$.
5: $\quad S' = S' \cup \{y_k^*\}$.
6: Form D^* the reordered matrix corresponding to S'.
7: Apply iVAT on D^* to obtain D'^* and produce $I(D'^*)$.
8: **return** S and D'^*.

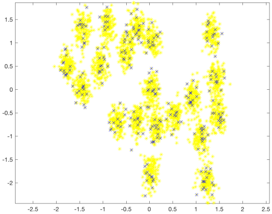
Theorem

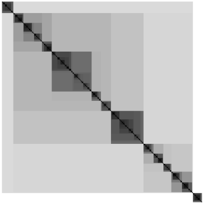
Sample obtained the algorithm is also a coreset of the given dataset T.

VAT results: compared with siVAT

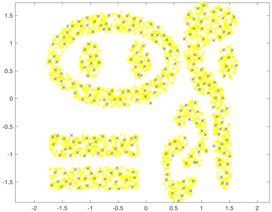


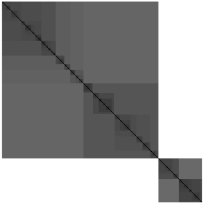


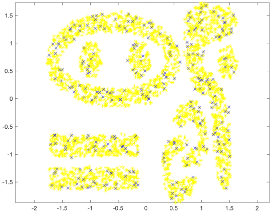


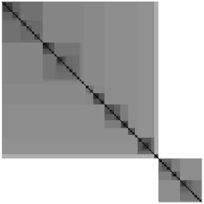


VAT results: deal with high complex structures







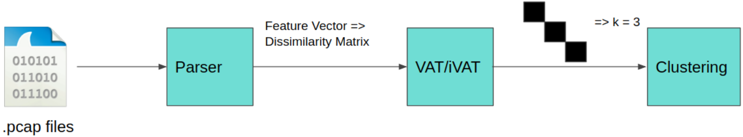


VAT for Bacnet datasets

A joint work with Prof. Fabio Massacci, Trentro University, Italy

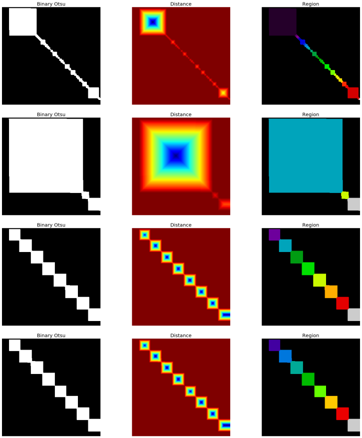
- BACnet: Building Automation and Control Networking Protocol

- Our proposed approach



VAT for Bacnet datasets

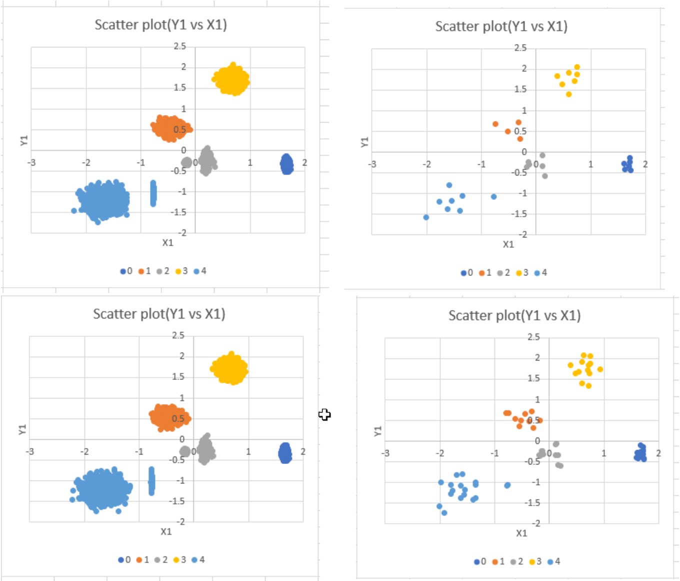
A joint work with Prof. Fabio Massacci, Trentro University, Italy



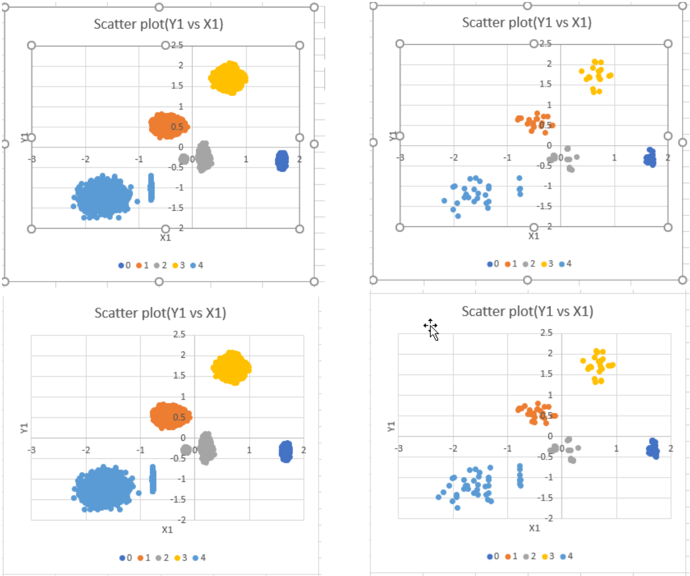
Binary image using otsu's threshold (left); the distance image from binary image (middle) and region image (right).

Streaming clustering: data processed with Spark

An example: results at t_0 and t_1



An example: results at t_2 and t_3



Clustering results: deal with streaming data

Datasets	Size	Cluster num.	Sample size	Whole dataset	Sample
A.set 1	3.000	20	55	23.24	**19.15**
A.set 2	5.250	35	61	43.56	**39.80**
A.set 3	7.500	50	59	54.52	**50.38**
FLAME	240	2	47	**18.70**	19
Birch-set 3	100000	100	143	518.03	**453**
JAIN	373	2	34	6.97	**6.85**
S.sets 1	5.000	15	52	21.43	**20.31**
S.sets 2	5.000	15	70	32.51	**31.24**
S.sets 3	5.000	15	70	32.14	**30.27**
S.sets 4	5.000	15	106	56.42	**54.24**
Dim 2	1.351	9	11	6.49	**7.49**
Unbalance	6500	8	25	14.47	**12.73**
D31	3100	31	62	21.64	**19.21**
G2-2-10	2048	10	23	19.05	**8.56**
G2-2-20	2048	20	43	19.50	**13.74**
G2-2-30	2048	30	76	**19.70**	19.98
G2-2-40	2048	40	89	**21.02**	22.89
Data1M-7	1000000	7	677	1255	**813**
Data1M-15	1000000	15	837	1542	**1027**
Data1M-55	1000000	55	2108	5400	**3342**
Data2M-77	2000000	77	2600	7800	**4500**

Outline

Summary

- A postprocessing task of the ProTraS is introduced to obtain a sample of the dataset such that
  - clusters in the sample are separated as much as possible,
  - while preserving the cluster structure of the whole dataset.

  $\rightarrow$ *obtain higher accuracy in VAT problem.*
- However,
  - ProTraS-based the sampling in our algorithm is also based on farthest-first traversal.
  - In the case of datasets with high noise or outliers, the algorithm might not be robust.
    - Maintain high representativeness points, while try to increase the inter-cluster distance.

Extension for a Clustering Algorithm

- Utilizing the proposed VAT algorithm to give an efficient clustering method dealing big data (with three features including Volume, Variety, and Velocity).
  - From VAT result on the sample set, try to obtain the clusters of the sample.
  - Generalize the result obtained on the sample to the whole dataset.

Outline

Problem

- Coreset for scaling applications in smart cities (with Bara and Mouzhi)
  - Improving the sample obtained by coreset.
  - Applying to scenarios in smart cities dealing with big datasets.
- VAT technique for anomaly detection in cybersecurity (discussing with Bacem)
  - Visualizing the cluster tendency for a streaming dataset.
  - Anomaly data points can be detected if they form a new dark block on the VAT image.

The End

Thank you for your attention.