Spatially Representative Online Big Data Sampling for Smart Cities

Isam Mashhour Al Jawarneh, Luca Foschini ,

Paolo Bellavista, Antonio Corradi, Rebecca Montanari

Department of Computer Science and Engineering - DISI
University of Bologna, Italy

{isam.aljawarneh3, luca.Foschini, paolo.bellavista, antonio.corradi,Rebecca.montanari}@unibo.it,

IEEE CAMAD 2020
14-16 September

Session 6 **Next Generation Networks**

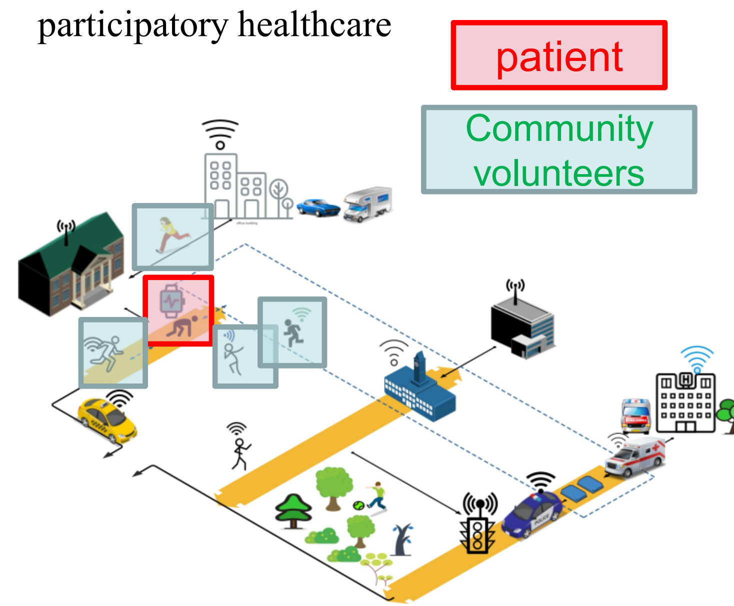
Agenda

- *Spatial Approximate Computing*: Background and Motivations
- *SpatialSPE*
    - SAOS spatial online sampling
    - Supported online queries
- *SpatialSPE* Deployment
    - Baseline system
    - Experimental setup
- Experimental Results
    - Extensive Microsoft Azure Spark cluster Test
- Conclusion

Motivating Application Scenario

A mixed-workload scenario requiring at least:

- **Traffic Light Controller**. Actuator decides to change lights consistently for ambulance to pass
- **Smart Real-time Pathfinder**. Interactive navigation map for ambulances and other vehicles
- **Real-time Community Detector**. Identify volunteers' communities in the surroundings of the patient



- **Primitive geospatial queries (expensive!)**
  - Proximity queries
  - Spatial join
  - Spatial clustering
  - Spatial geo-statistics.
  - *k*-Nearest Neighborhoods)

- **Data arrives fast during peak hours**
- **Exceeds the capacity of ingestion and processing systems**

- Spatial Approximate Query Processing (SAQP) is the key.

QoS Tension for achieving SLAs

Spatial (Approximate) Query Processing (S(A)QP)

Tension

latency

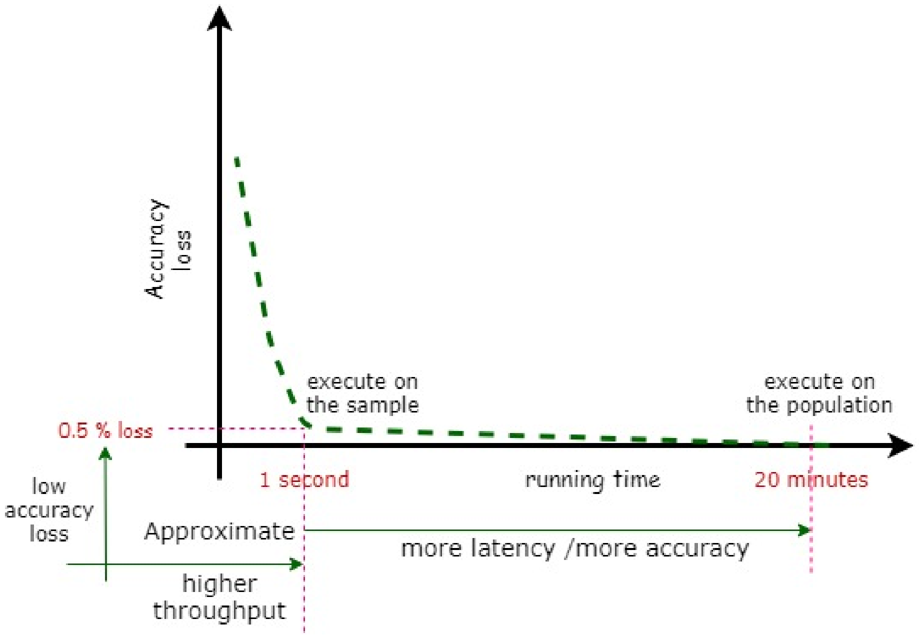
Throughput

Accuracy

Spatial Approximate Query Processing (SAQP)

- Exactness is not necessary for decision making in smart cities!

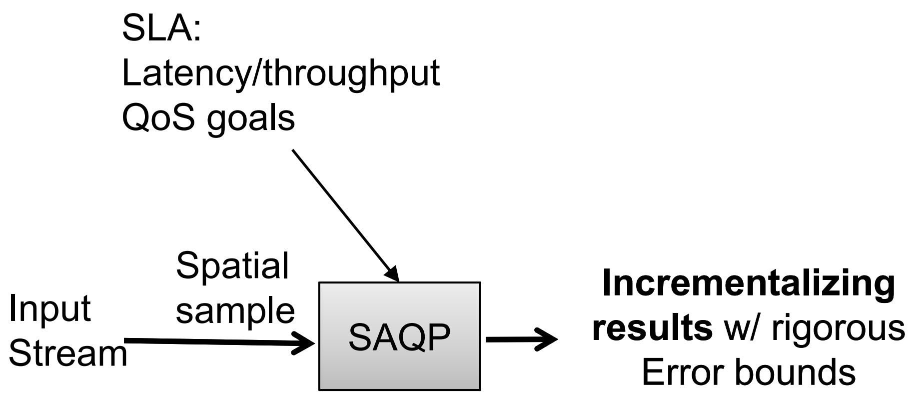
✓ After 1 second, we obtain a 99.5 accurate early result, which is satisfactory for decision making, which then makes the final exact result not needed.



Spatial Sampling

- **SAQP** employs a spatial sampling design aiming at resolving gracefully the tension between low-latency and high-accuracy QoS goals.
- ***Spatial sampling***. Selecting a miniscule version of the population to compute geo-statistics: mean, range, total.
  - Based on the fact that decision makers are can withstand a tiny loss in ***error-bounded accuracy*** in exchange for a plausible ***latency gain***.
  - In streaming settings, data keeps arriving, the 'population' metaphor vanishes.

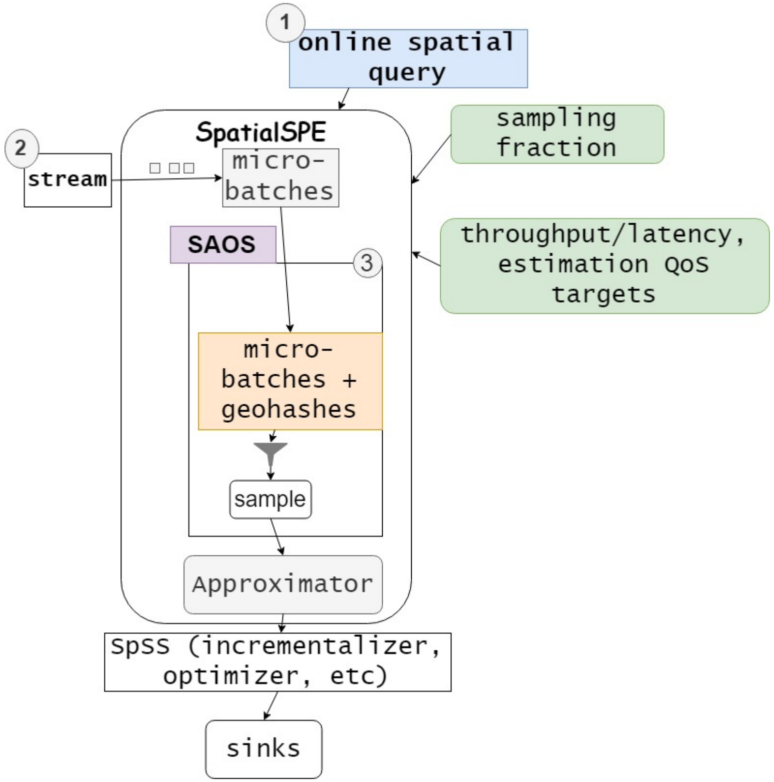
Relying on samples for computing geo-stats



Plain SpatialSPE

- Nearby spatial objects share a pairwise relationship
  - ***spatially well-balanced representative samples*** → are known to yield better results for geo-stats (average, median, etc.) in terms of accuracy.
- ***Example online spatial query.*** "what is the average trip distance travelled by taxis from each neighbourhood in the city of Rome, Italy"
- By sampling the same fraction from each geohash, we **approximately** guarantee that each neighborhood (stratum in statistics parlance) is fairly represented
- Continuous results are updated by ***incrementalization.***

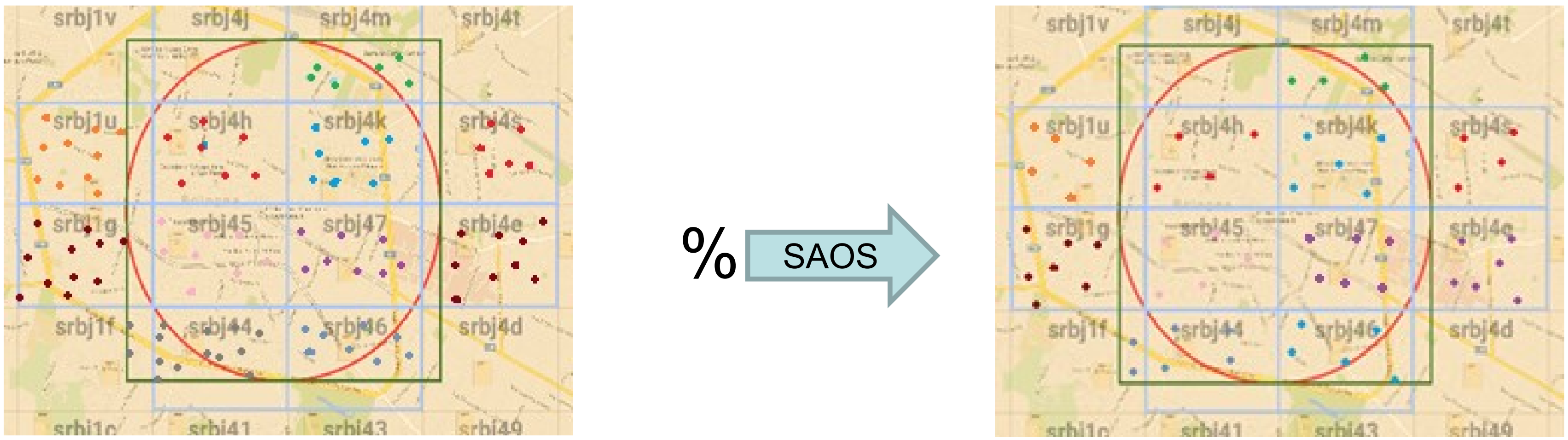
SpatialSPE overview



Spatial Aware Online Sampling (SAOS): overview

- Nearby points share the same geohash prefixes, thus reducing the two-dimensional point representations to one-dimensional string ordering.
- **Geohash** indexing. An ordering (string representation) imposed on grid surface earth planar representation.

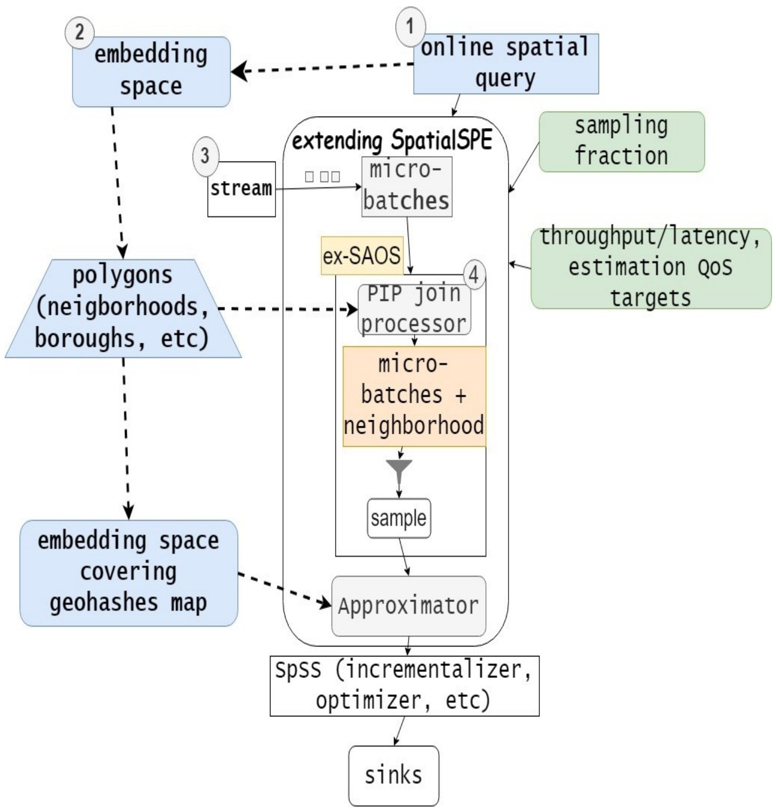
Granular



- Nearby points share the same geohash prefixes
- Only the ‘filter’ stage of the ‘filter-and-refine’!
- **SAOS** focuses on **SDL preservation,** but with ‘**false positives**’
- ‘**False positives**’ are those tuples that have the same geohash, but do not belong to the same neighborhood

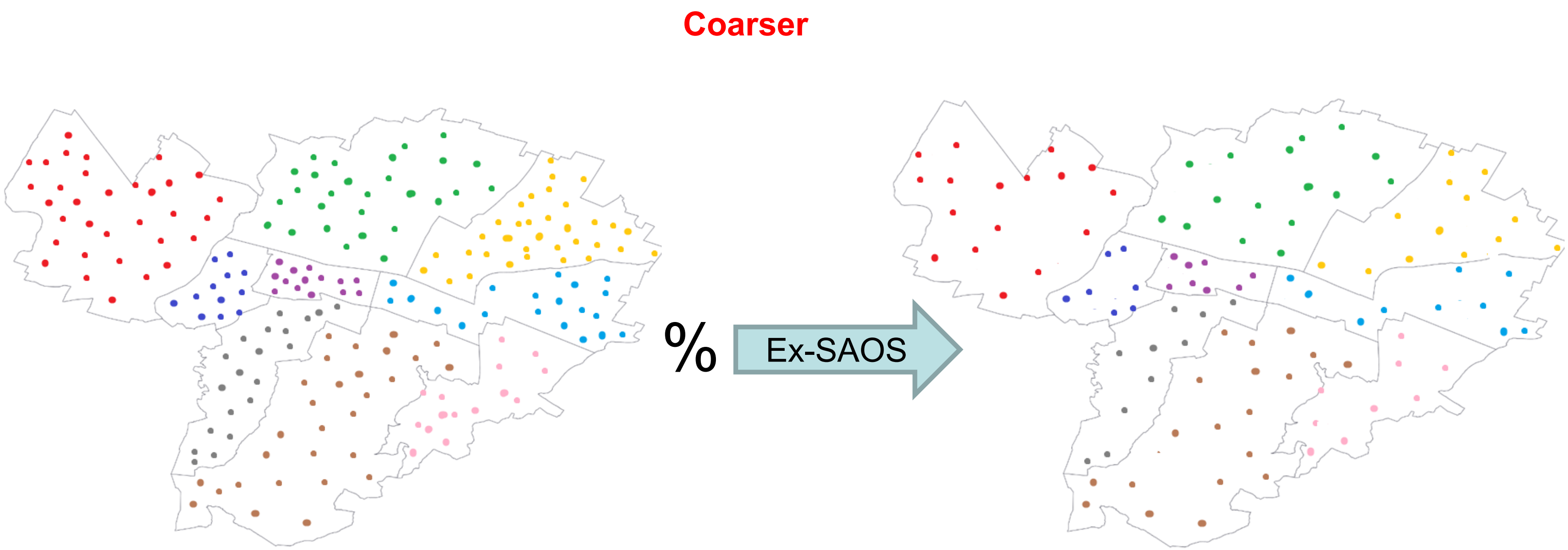
Incorporating Ex-SAOS in SpatialSPE

- Applying ‘filter-and-refine’ to solve the PIP test before sampling.
- Discarding ‘false positives’.
- We exactly sample **same fractions** from each neighbourhood (borough, district, etc.,)
- Yields more accurate results.



Extended SpatialSPE overview

Extended Spatial Aware Online Sampling (SAOS): overview



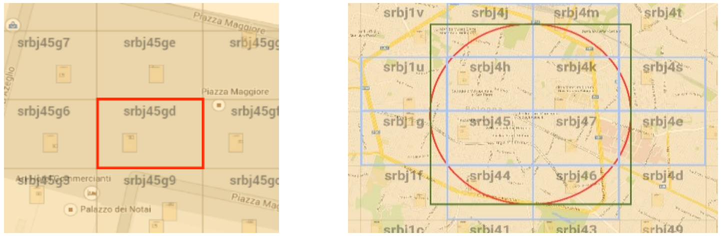
- Applying all stages of ‘filter-and-refine’!
- **Ex-SAOS** focuses on **SDL preservation,** without ‘**false positives**’
- More accurate.

Plain Spatial Aware Online Sampling (SAOS)

- A ***hybridization*** between ***z-order curves*** (geohash) and ***simple probability sampling*** (within each grid cell).
- does not require a pre-knowledge of the streaming statistics, it otherwise depends on ***incrementalization***.

Algorithm 2: Spatial-Aware Online Sampling (SAOS)

SAOS (micro-batch-tuples, samplingMap, samplingFraction, seed)
r = random(seed)
S ← ∅
Foreach tuple **in** micro-batch-tuples **do**
  geohash ← geocode (tuple)
  //get the sampling fraction for this geohash key = $fraction_i$*, or zero if not present.*
  $fraction_i$ ← samplingMap.getOrElse(geohash,0.0)
  //toss a coin for selecting items belonging to each geohash from the current batch interval
  If (P (r < $fraction_i$))
  S.put(tuple)
  End
End



heuristic overview

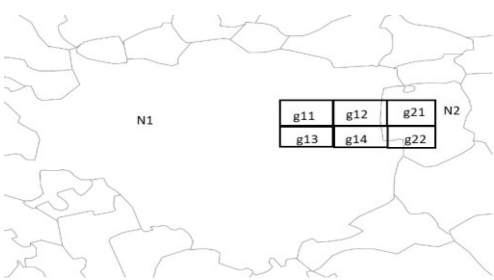
- **Geohash** indexing. An ordering (string representation) imposed on grid surface earth planar representation.
- Nearby points share the same geohash prefixes, thus reducing the two-dimensional point representations to one-dimensional string ordering.

Extended *Spatial Aware Online Sampling (EX-SAOS)*

- Sampling proportionally balanced tuples from each administrative division (neighborhood, district, borough, etc.,) independently.
- 'Filter-and-refine' spatial join for resolving the Point-in-Polygon (PIP).

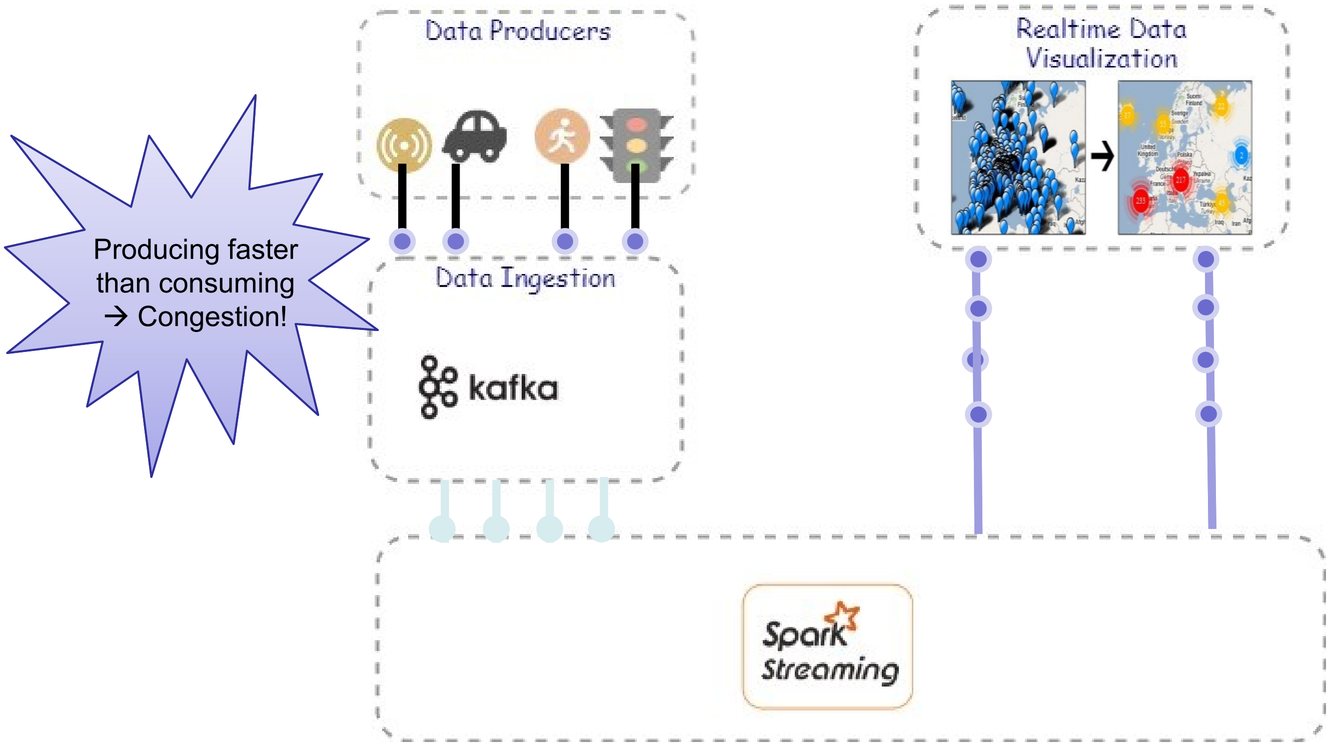
Algorithm 1: Extended- Spatial-Aware Online Sampling (ex-SAOS)

```
ex-SAOS (tuples_i, samplingMap, coverGeo, sampFraction, seed)
r = rand(seed), sample ← {}
//perform inner join on geohash
joinResult = tuples_i.join(coverGeo)
Foreach tuple t in joinResult do
//return the polygon to which this tuple belongs
  polygon ← getPolygon (t)
  //get sampling fraction for this polygon key = fraction_i, or zero
  fraction_i ← samplingMap.getOrElse(polygon,0.0)
  //toss a coin selecting items from each polygon in current batch
  If (P (r < fractioni)) S.put(tuple)
return S
```

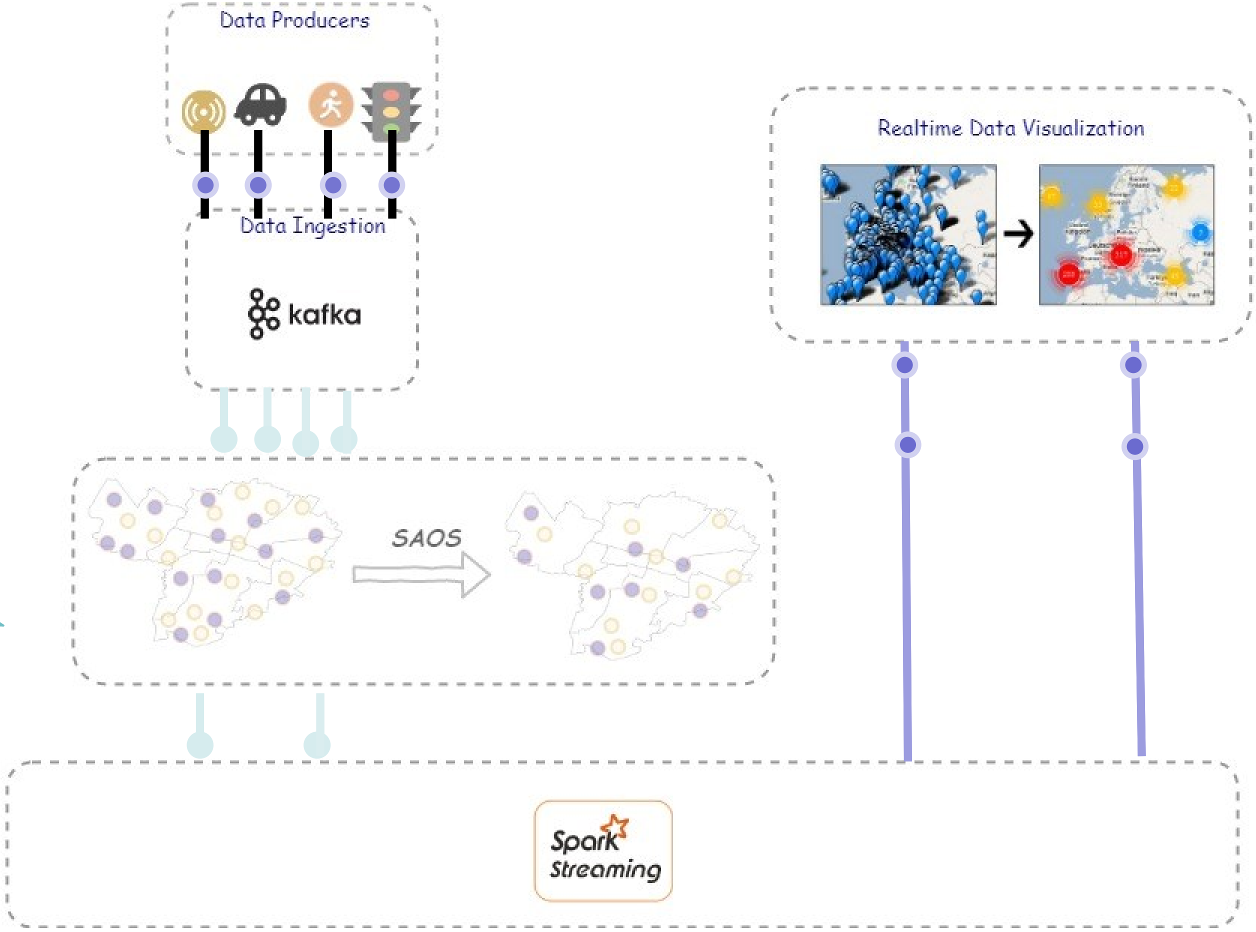


heuristic overview

Typical pipeline architecture w/o SAOS or Ex-SAOS



The improved architecture w/ Ex-SAOS



SQL-alike continuous geospatial query

Same serving speed, but serving less data!

Supported Queries

Single queries

"average trip distance travelled by taxis in Rome, Italy during the last three months"

$$\bar{Y}_{Ex-SAOS} = \frac{\hat{t}_{Ex-SAOS}}{\mathrm{N}} = \sum_{i=1}^{I} \frac{N_i}{N} \bar{y}_i$$

$$RE = \frac{z_{\alpha/2} SE(\bar{Y}_{Ex-SAOS})}{\bar{Y}_{Ex-SAOS}}$$

Online continuous query

Incremental geo-stats

error bounds (e.g., relative error RE) and spearman's rho

Stateful aggregation

"top-3 boroughs in Bologna city in Italy where people tend to check out shared bikes".

$spatial\ incremental$
Top-N online aggregation function

$$\rho_{rg} = \frac{cov(rank_{nos}, rank_{samp})}{\sigma_{rank_{nos}} \sigma_{rank_{samp}}}$$

No pre-knowledge on the streaming geo-statistics is required, we depends on ***incrementalization***

Baselines

Spark Streaming

SRS, Stratified Sampling

Spatial-awareness is not supported

Spark Structures Streaming

Sampling is not supported

Depending on SRS for selecting a miniscule sample from a highly skewed spatial dataset yields highly inaccurate results.
This is so because it tends to overlook regions while underlooking other regions.

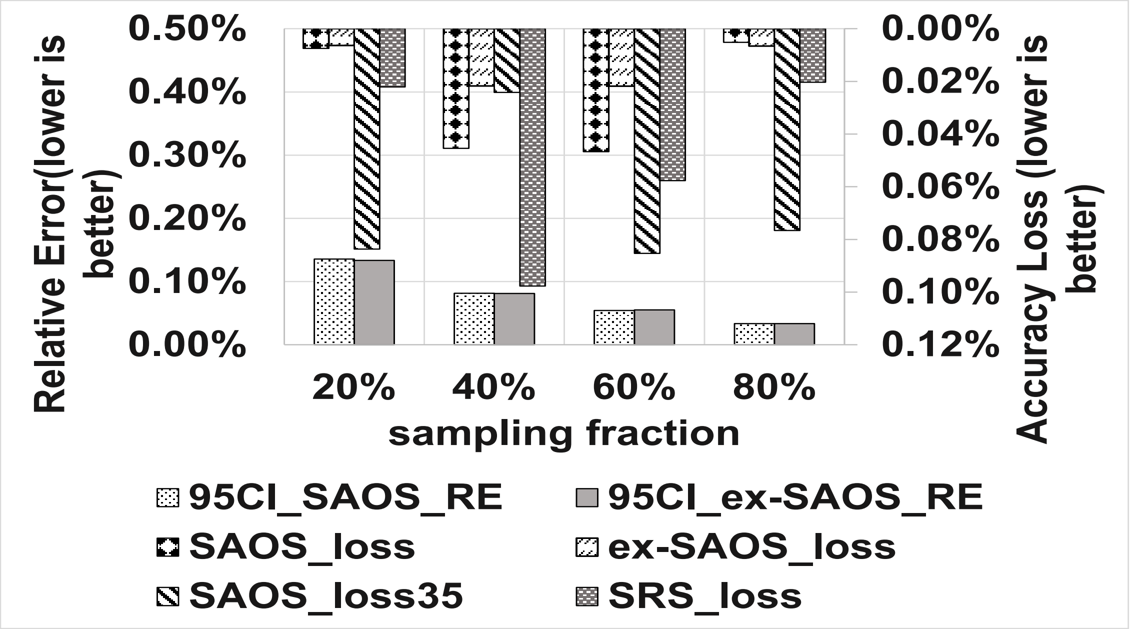
Baseline System

- Our baseline is a sampling method that we have designed in previous work (termed as SAOS).
- We have also incorporated transparently an SRS version with SpSS and compared the novel method ex-SAOS with it.

Experimental setup

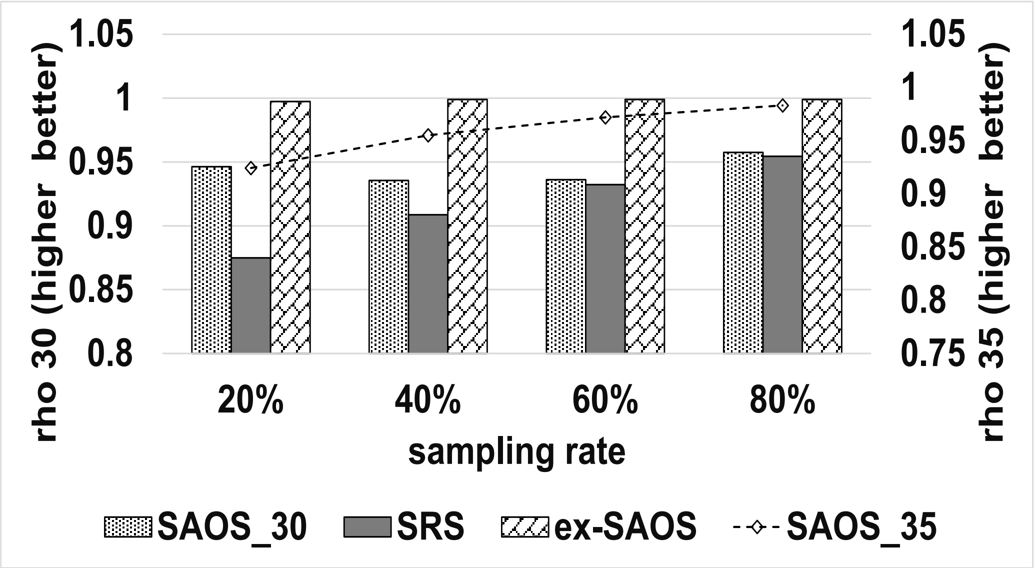
- Evaluation metrics
  - Sampling fraction vs accuracy
  - Sampling fraction vs rho
- Testbed
  - Cluster: 6 nodes (Microsoft Azure HDInsight Cluster)
  - Datasets:
    - NY City taxicab trips datasets (cohort of six months dataset (around nine million units)

Sampling fraction vs Accuracy



- We define the accuracy loss as accLoss = |estimatedMean – trueMean| / trueMean.
- ex-SAOS outperforms SpSS-based SRS in addition to SAOS for all geohash values for all measures, accuracy loss and relative error.
- Ex-SAOS has lower accuracy loss compared to all geohash precisions applied to SAOS (30 & 35)

Spearman's rho vs Sampling fraction



- rho is a measure for statistical dependency between the ranking of two variables in a dataset.
- Ranking precision of ex-SAOS outperforms those for SAOS(with all precisions 30 & 35) in addition to SRS.

Concluding remarks

- Most interesting locational intelligence queries are required during high-pace data streaming arrival rates, where SPEs can not withstand the speed!
- Spatial sampling based on stratified designs is proven to yield more plausible geo-stats.
- We have extended a plain SAOS by enabling a spatial sampling design on a granular level that causes results in more accuracy.

Q&A and Contacts

Thanks for your attention!

Questions time...

Isam Al Jawarneh
Email: isam.aljawarneh3@unibo.it

Luca Foschini
Email: luca.foschini@unibo.it

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI INFORMATICA - SCIENZA E INGEGNERIA