جـامعـــة الشـــارقـة
UNIVERSITY OF SHARJAH

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

Efficient Generation of Approximate Region-based Geo-maps from Big Geotagged Data

Dr. Isam Mashhour Al Jawarneh, **Dr. Luca Foschini**, Prof. Antonio Corradi

IEEE CAMAD 2023

11/2023

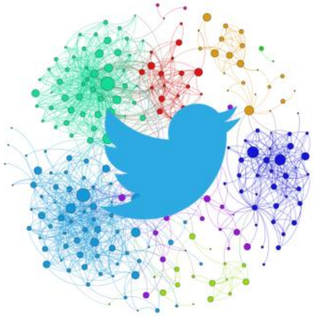
Outline

- ➢ Introduction
  - o Motivating scenario
  - o Spatial data challenges & requirements
- ➢ A method for generating region-based approximate geo-maps
  - o Overview
  - o ApproxGeoViz
- ➢ Results and Discussion
  - o Deployment: baselines & testing setup
  - o ApproxGeoViz Vs. baseline
- ➢ Summary & future research

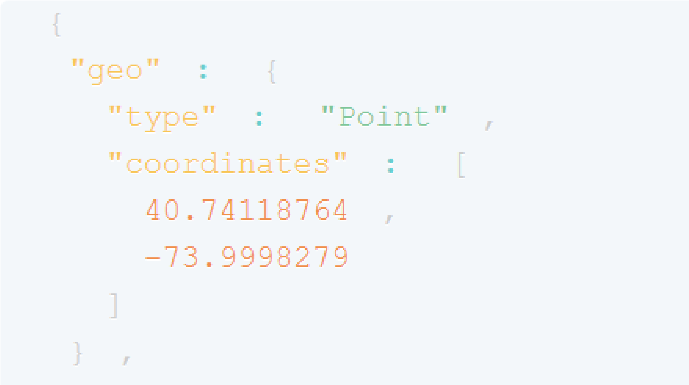
Big data examples

- **YouTube** : Several petabytes (~**350 PB** of data in 2019)
- **500-700** million **tweets** a day,
  - which adds up to roughly **12 terabytes** of data every 24 hours.
- **Facebook**
  - on the verge of **500** daily **terabytes**,

Source: Forbes



Tweet with exact location



```
{
 "geo"  :   {
   "type"  :   "Point"  ,
   "coordinates"  :   [
     40.74118764  ,
     -73.9998279
   ]
 }  ,
```

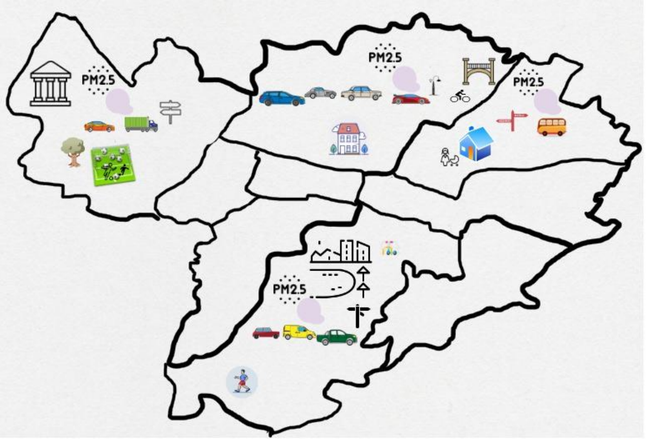
- Most data (**>60%**) is **geo-referenced**!

Spatial Data-intensive applications

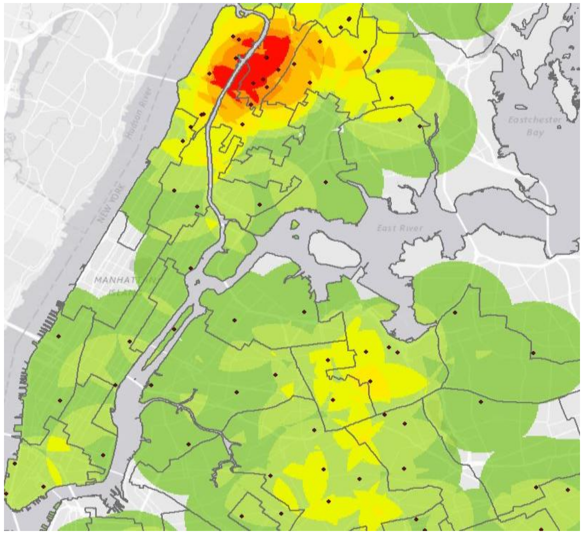
- Spatial Data is the primary **challenge**
  - **Volume (size)**,
  - **Complexity**,
  - **Speed** of arrival & **change** (**uncertainty**)

Motivating scenario

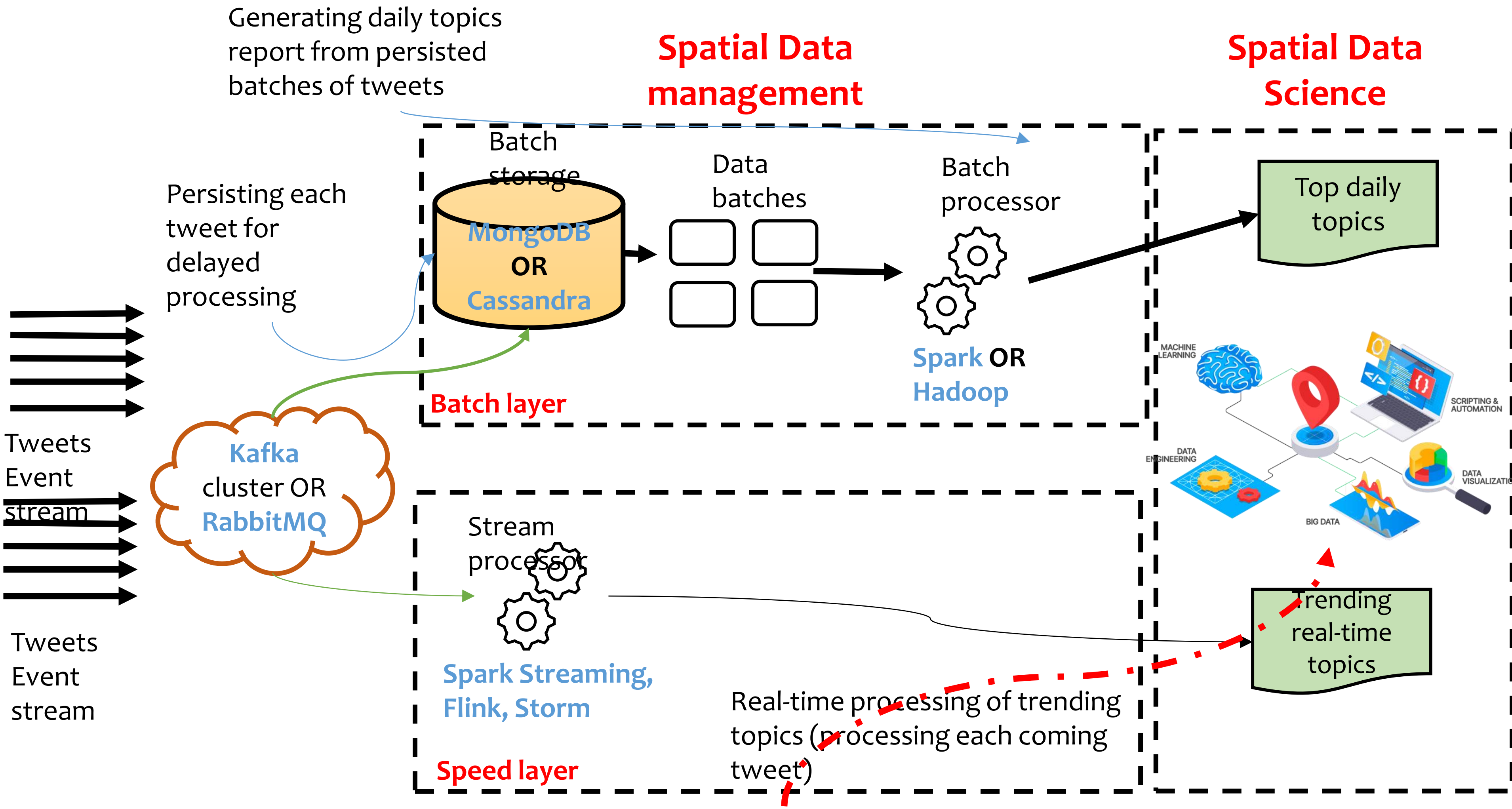
- Billions of GPS-enabled handheld devices collect massive data amounts
- Data is subjected to Exploratory Spatial Data Analytics (**ESDA**)
  - generating geo-maps (e.g., **region-based** maps such as choropleth)



- **Geospatial** aggregation
  - Air pollutants **density** in each **zone**,
  - **Autocorrelation** between nearness and pollution



Lambda: a reference architecture



Generating daily topics report from persisted batches of tweets

Spatial Data management

Spatial Data Science

Persisting each tweet for delayed processing

Batch storage

MongoDB OR Cassandra

Data batches

Batch processor

Spark OR Hadoop

Top daily topics

Batch layer

Tweets Event stream

Kafka cluster OR RabbitMQ

Tweets Event stream

Stream processor

Spark Streaming, Flink, Storm

Speed layer

Trending real-time topics

Real-time processing of trending topics (processing each coming tweet)

Big data **geo-visualization** is an integral part of the pipeline

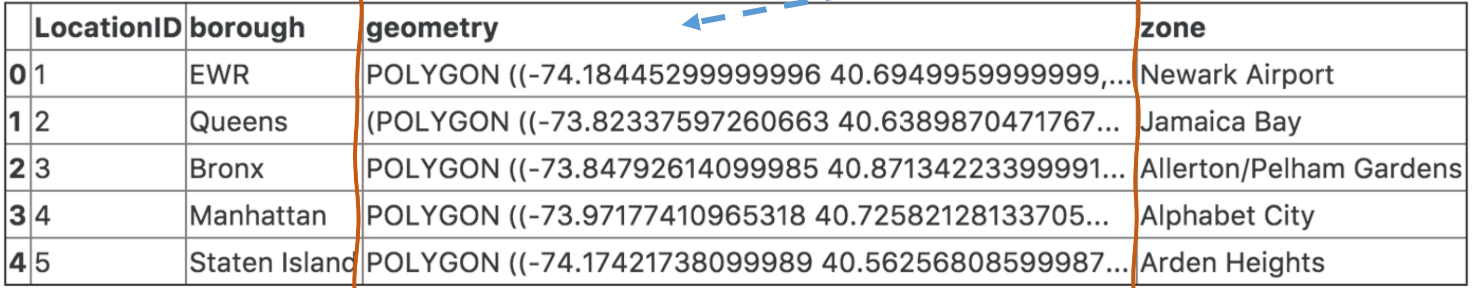
Outline

- Geospatial big data analytics: Background and Motivating scenario
  - Motivating scenario
  - Spatial data challenges & requirements
- A method for generating region-based approximate geo-maps
  - Overview
  - ApproxGeoViz
- Results and Discussion
  - Deployment: baselines & testing setup
  - ApproxGeoViz Vs. baseline
- Summary & future research

Spatial data analytics challenges

Shapefile, NYC

Polygons



	LocationID	borough	geometry	zone
0	1	EWR	POLYGON ((-74.18445299999996 40.6949959999999,...	Newark Airport
1	2	Queens	(POLYGON ((-73.82337597260663 40.6389870471767...	Jamaica Bay
2	3	Bronx	POLYGON ((-73.84792614099985 40.87134223399991...	Allerton/Pelham Gardens
3	4	Manhattan	POLYGON ((-73.97177410965318 40.72582128133705...	Alphabet City
4	5	Staten Island	POLYGON ((-74.17421738099989 40.56256808599987...	Arden Heights

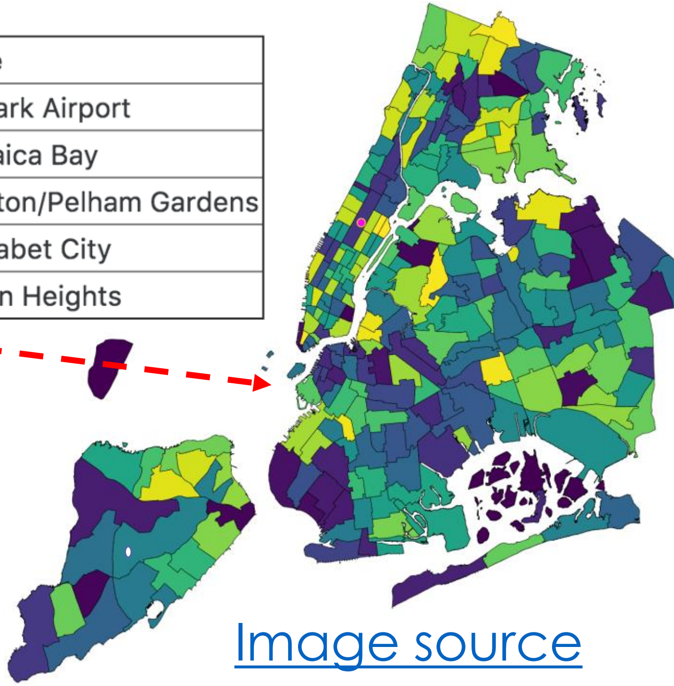
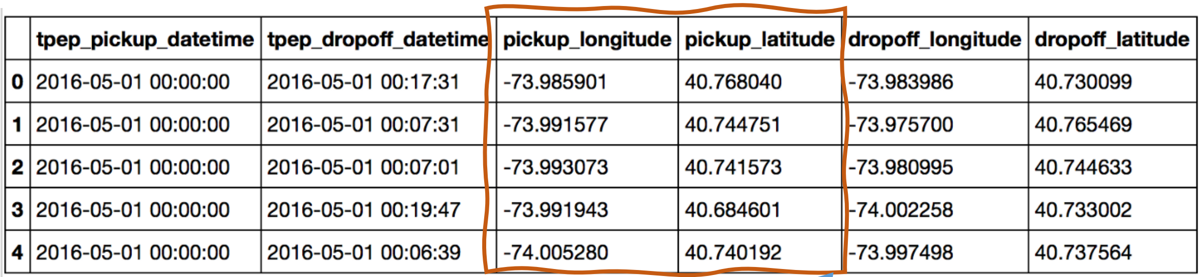


Image source

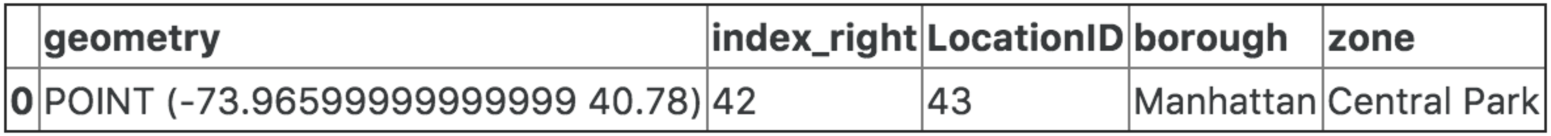
taxi dataset



	tpep_pickup_datetime	tpep_dropoff_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
0	2016-05-01 00:00:00	2016-05-01 00:17:31	-73.985901	40.768040	-73.983986	40.730099
1	2016-05-01 00:00:00	2016-05-01 00:07:31	-73.991577	40.744751	-73.975700	40.765469
2	2016-05-01 00:00:00	2016-05-01 00:07:01	-73.993073	40.741573	-73.980995	40.744633
3	2016-05-01 00:00:00	2016-05-01 00:19:47	-73.991943	40.684601	-74.002258	40.733002
4	2016-05-01 00:00:00	2016-05-01 00:06:39	-74.005280	40.740192	-73.997498	40.737564

Points (parametrized) Projected Coordinate System (PCS)

assigning trips pickups to city zones (districts) is an example of a **spatial join (expensive)**



	geometry	index_right	LocationID	borough	zone
0	POINT (-73.96599999999999 40.78)	42	43	Manhattan	Central Park

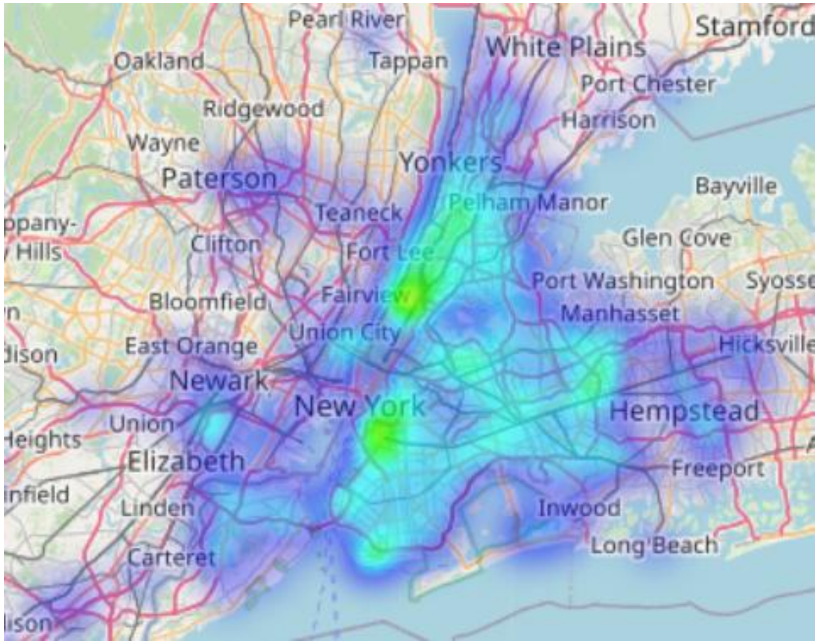
Outline

- ➢ Geospatial big data analytics: Background and Motivating scenario
  - o Motivating scenario
  - o Spatial data challenges & requirements
- ➢ A method for generating region-based approximate geo-maps
  - o Overview
  - o ApproxGeoViz
- ➢ Results and Discussion
  - o Deployment: baselines & testing setup
  - o ApproxGeoViz Vs. baseline
- ➢ Summary & future research

Geo-visualization Process

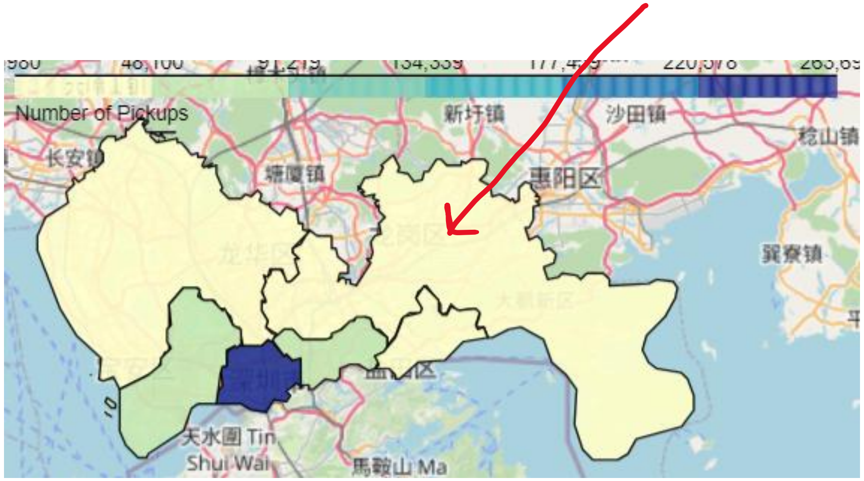
- Geospatial data **processing**
  - e.g., for generating a choropleth map, data needs to be aggregated into clusters
  - geospatial stateful aggregation queries such as grouping by
- Geospatial data **visualization**
  - rasterizing the vector data and rendering it into maps that viewed on screen

Geo-visualization examples



NYC taxi pickups heatmap

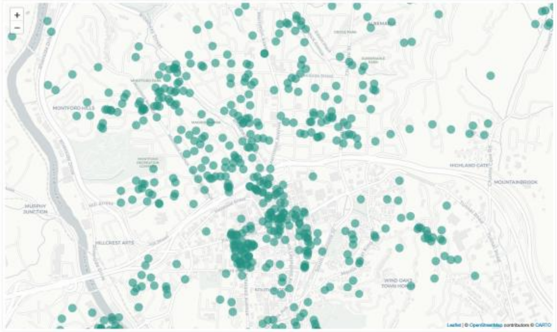
predefined tessellation



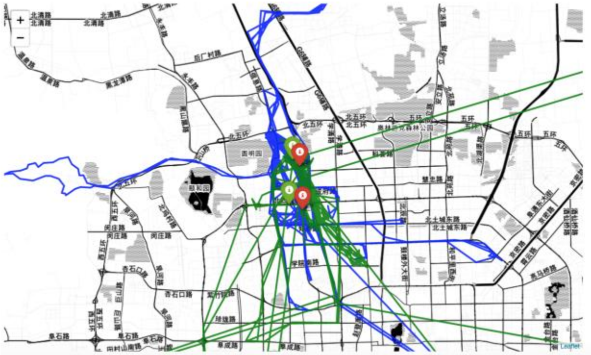
Shenzhen (China) electric taxi pickup choropleth

Approaches for visualizing georeferenced data

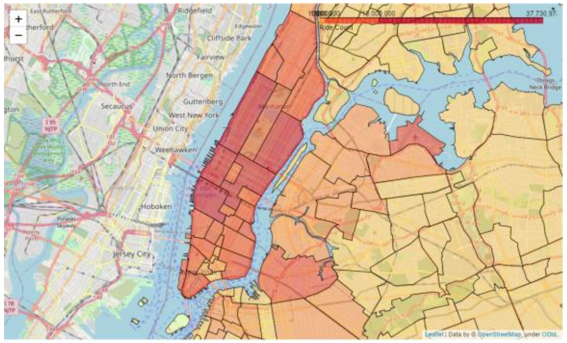
- **Point-based**
  - plot individual points on geographical maps such as Point-of-Interest (POI)



- **line-based**
  - time-series trajectory visualization of spatial data



- **region-based (costliest)**
  - tessellating geographic regions into grid cells, then, grouping data by region-based aggregations
  - e.g., **Choropleth** maps generation



Challenges in generating **region-based** maps

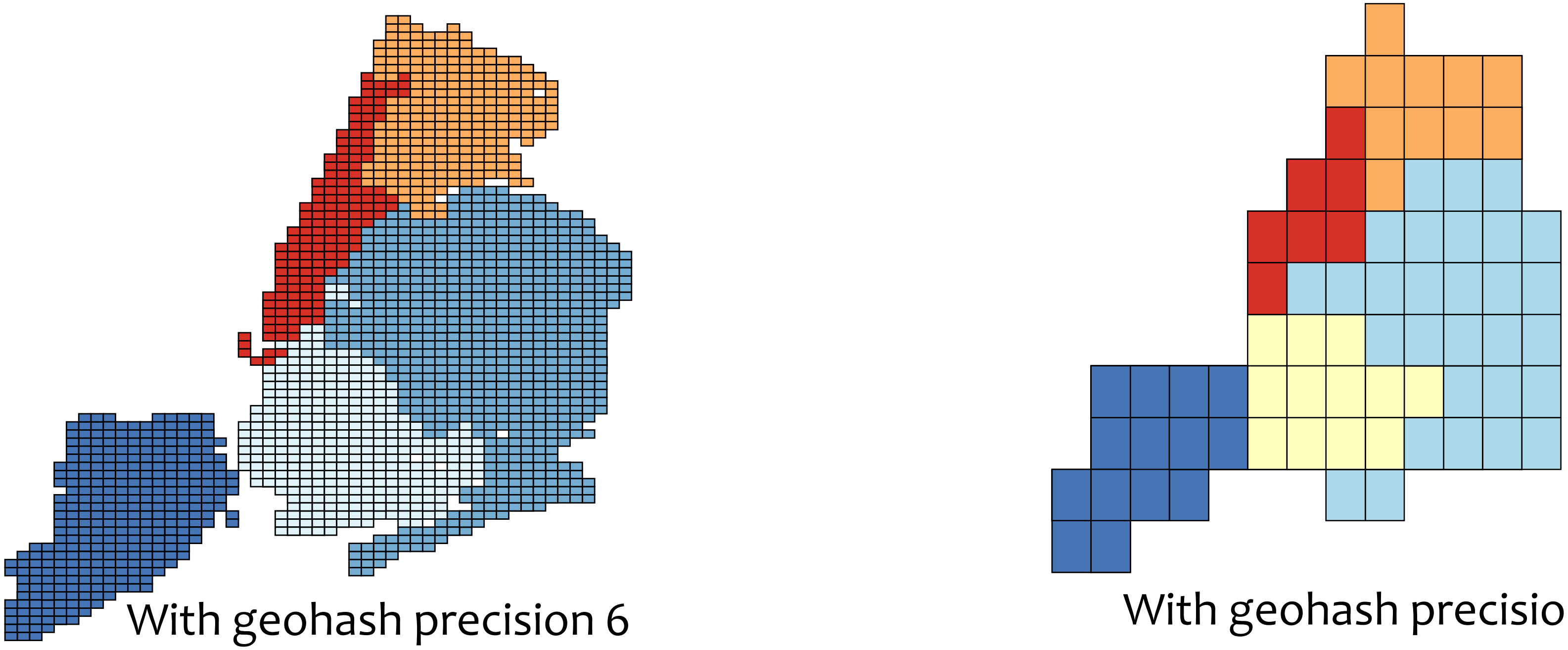
- Region-based geo-maps require stateful data **aggregation**
  - Computationally **expensive** in real data stream settings
    - Georeferenced data is typically **parametrized**
    - Brining them into their original forms, is a kind of **geospatial join** (computationally **costly**)
  - Out-of-service during spikes in arrival rates
- Geospatial data **preprocessing** (including aggregation) is the **dominating** component for generating geospatial region-based maps

Coping up with geo-data loads

- **Scalability**
    - Hardware scalability. **Overprovisioning** resources
    - Scaling **up/out**
- **Approximate Query Processing** (**AQP**). Data reduction
    - **Spatial** Approximate Query Processing (**SAQP**)
    - e.g., load shedding and geospatial **sampling**

Our focus!

SAQP: Geohash tessellation



With geohash precision 6

With geohash precision 5

Geohash tessellation for NYC city, USA

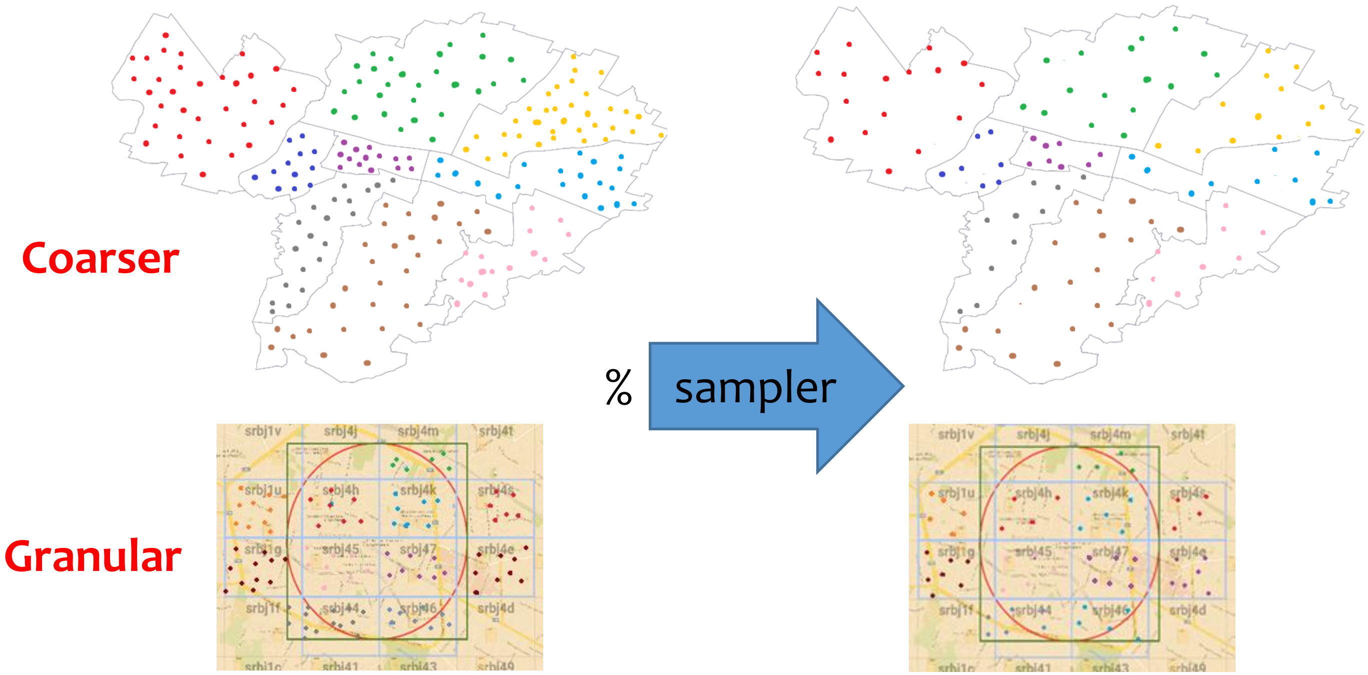
- Can be used for **stratified-like** sampling
  - Captures the reality
  - Each geohash is a **strata**
  - All geohash covering the area are **stratum**

Outline

- ➢ Geospatial big data analytics: Background and Motivating scenario
  - ○ Motivating scenario
  - ○ Spatial data challenges & requirements
- ➢ A method for generating region-based approximate geo-maps
  - ○ Overview
  - ○ ApproxGeoViz
- ➢ Results and Discussion
  - ○ Deployment: baselines & testing setup
  - ○ ApproxGeoViz Vs. baseline
- ➢ Summary & future research

Geohash-based geospatial stratified-like sampler

A **design** for generating **region-based** **approximate** geo-maps from voluminous geotagged data



Coarser

Granular

% sampler

- Nearby points share the same geohash prefixes
- **Stratified-like sampler** focuses on **spatial co-locality preservation**
  - yield more accurate results than random sampling counterparts

ApproxGeoViz: Geospatial Visualization at Scale with QoS Guarantees

geo-map **renderer** (visualizer)

Five components

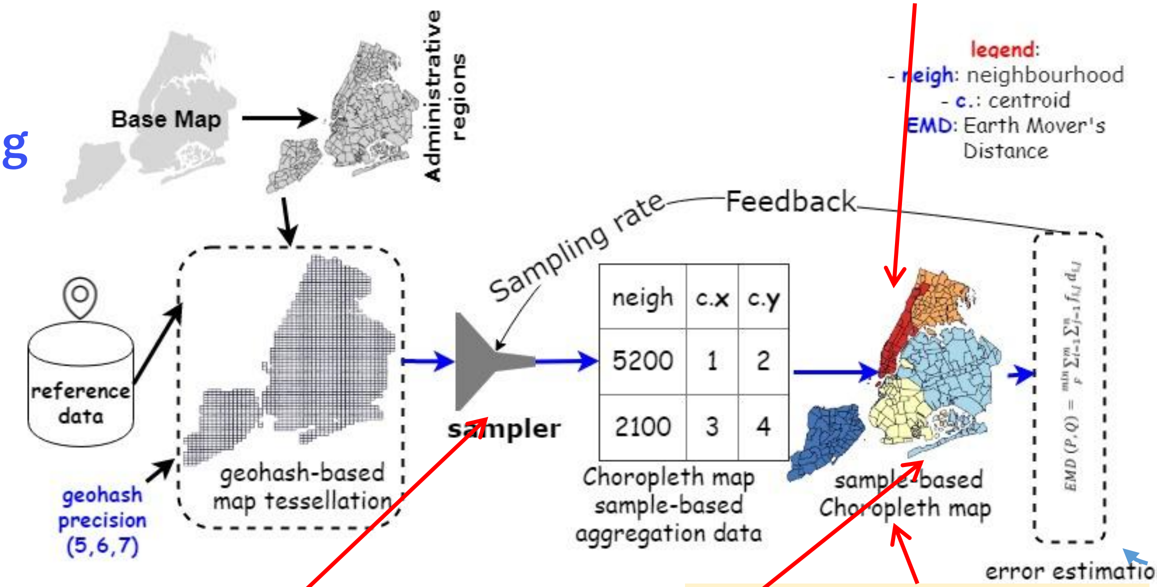
(1) Geospatial data **modelling** and **representation**

(2) Stratified-like geospatial **sampler**

(3) Region-based geo-map **proxy generator**

(4) Geo-map **renderer** (visualizer), and

(5) QoS **controller**



Base Map
Administrative regions
reference data
geohash precision (5,6,7)
geohash-based map tessellation
Sampling rate
Feedback
sampler
Choropleth map sample-based aggregation data
sample-based Choropleth map
legend:
- neigh: neighbourhood
- c.: centroid
EMD: Earth Mover's Distance
$EMD(P,Q) = \frac{min}{F} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}$
error estimation

neigh	c.x	c.y
5200	1	2
2100	3	4

Geohash-based stratified-like

aggregates data in all geohashes covering each polygon individually

EMD is a distance metric for measuring similarity between two data distributions or densities

Proxy: a compact representation of aggregated vector data as a matrix

$$EMD\ (P, Q) = \frac{min}{F} \sum_{i=1}^{m} \sum_{j=1}^{n} flow_{i,j}\ dist_{i,j}$$

Outline

- ➢ Geospatial on big data analytics: Background and Motivating scenario
  - o Motivating scenario
  - o Spatial data challenges & requirements
- ➢ A method for generating region-based approximate geo-maps
  - o Overview
  - o ApproxGeoViz
- ➢ Results and Discussion
  - o Deployment: baselines & testing setup
  - o ApproxGeoViz Vs. baseline
- ➢ Summary & future research

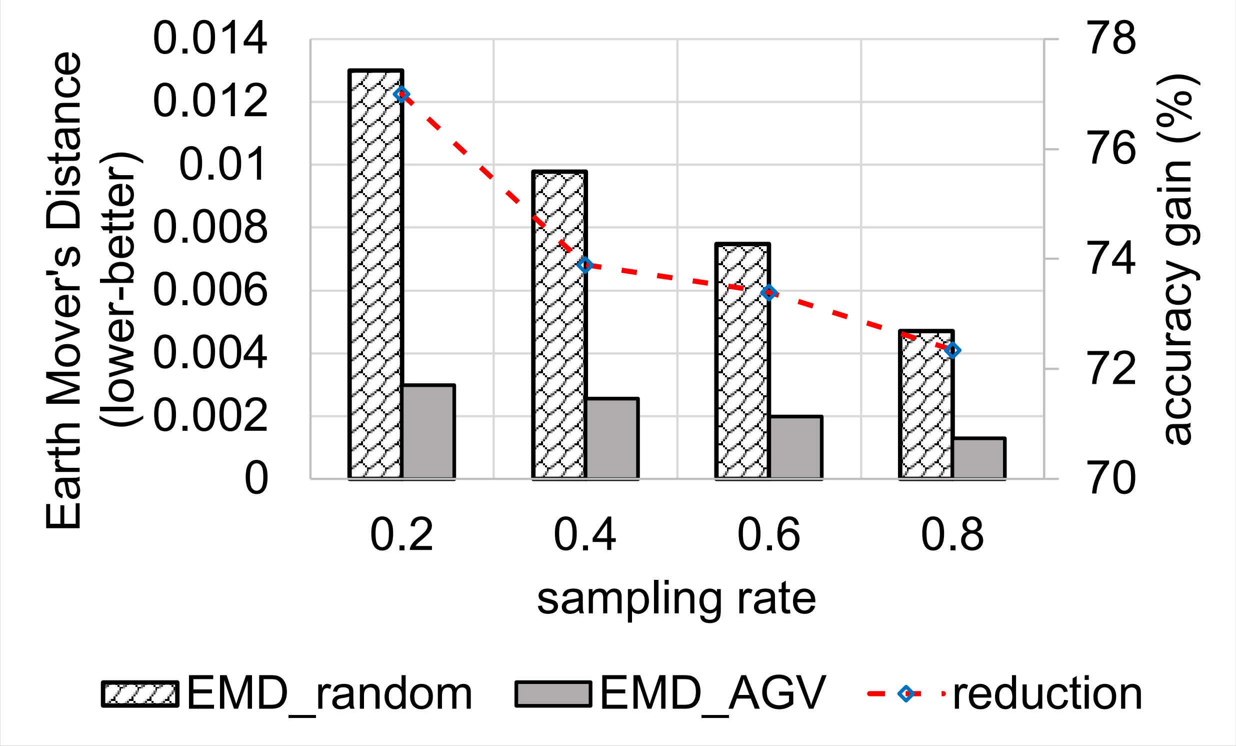
Experimental setup

- **Evaluation metrics**
  - Earth Mover’s Distance (EMD) distance measurement
- **Baselines**
  - Plain region-based geo-map generator with random sampler
- **Testbed**
  - We have deployed **ApproxGeoViz** on a Microsoft Azure virtual machine hosting Python
  - **Datasets**
    - Vehicle mobility dataset
      - Uber pickup dataset from the city of San Francisco in USA
      - anonymized GPS coordinates (longitudes/latitudes) of Uber taxi trips forming around one million and 85k tuples

Outline

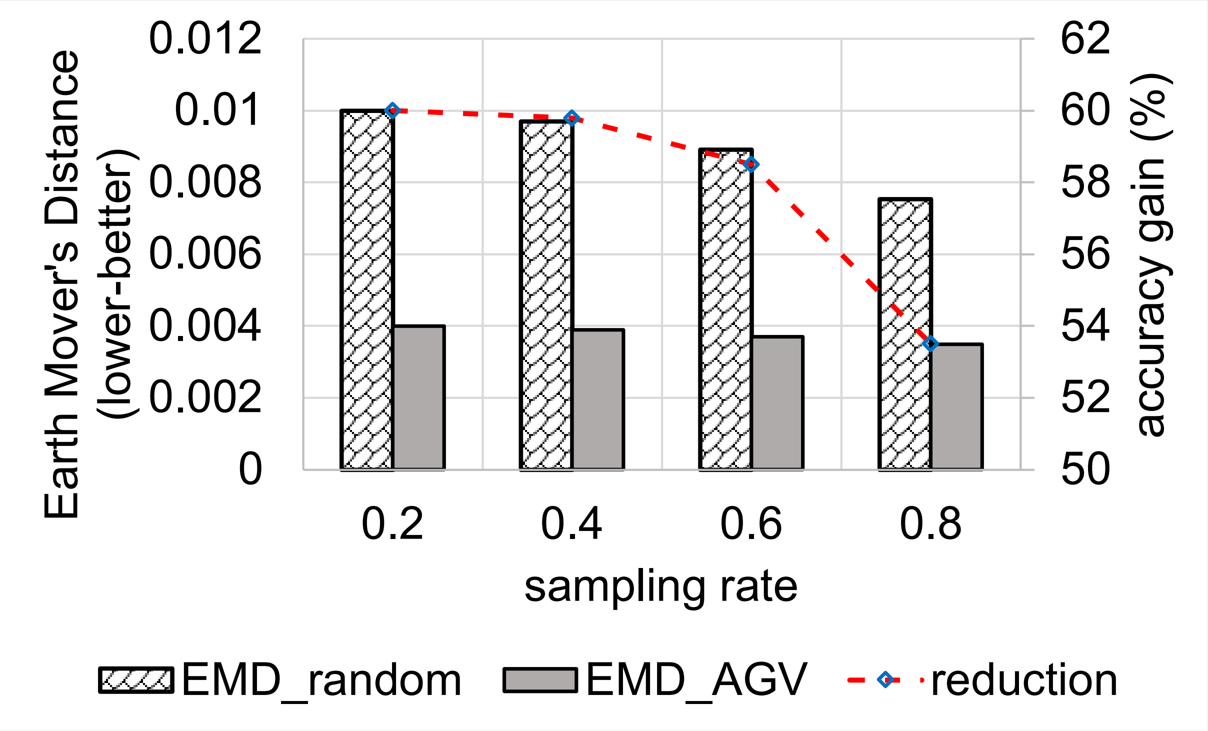
- ➢ Geospatial big data analytics: Background and Motivating scenario
  - ○ Motivating scenario
  - ○ Spatial data challenges & requirements
- ➢ A method for generating region-based approximate geo-maps
  - ○ Overview
  - ○ ApproxGeoViz
- ➢ Results and Discussion
  - ○ Deployment: baselines & testing setup
  - ○ ApproxGeoViz Vs. baseline
- ➢ Summary & future research

EMD of ApproxGeoViz Vs. baselines: geohash 6



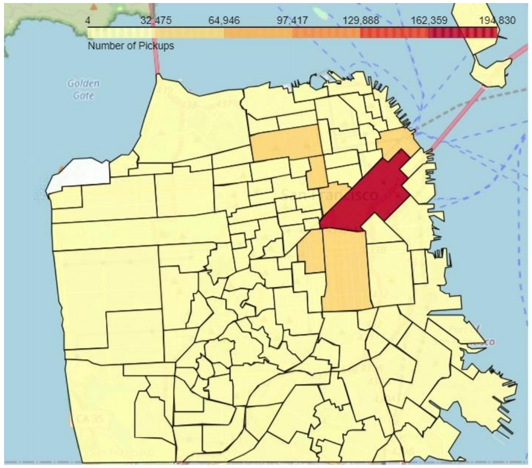
- **Varying** the geohash precision and sampling rate and
- **computing** EMD to test performance of system by applying both samplers (stratified-like Vs. baseline)
- On average, an accuracy **gain** that roughly equals to 74%

EMD of **ApproxGeoViz** Vs. baselines: geohash **5**

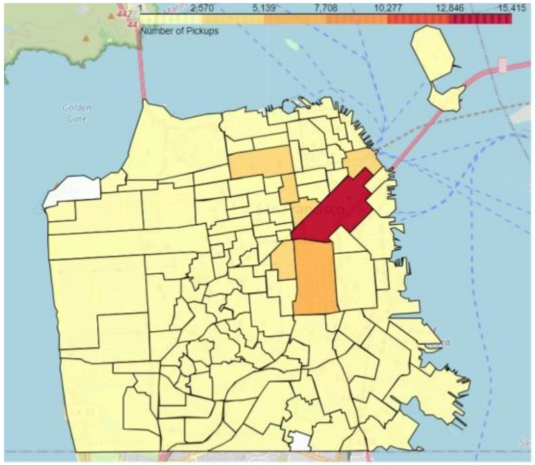


- **Varying** the geohash precision and sampling rate and
- **computing** EMD to test performance of system by applying both samplers (stratified-like Vs. baseline)
- On average, an accuracy **gain** that roughly equals to 57.9%
- less than the case of geohash precision 6

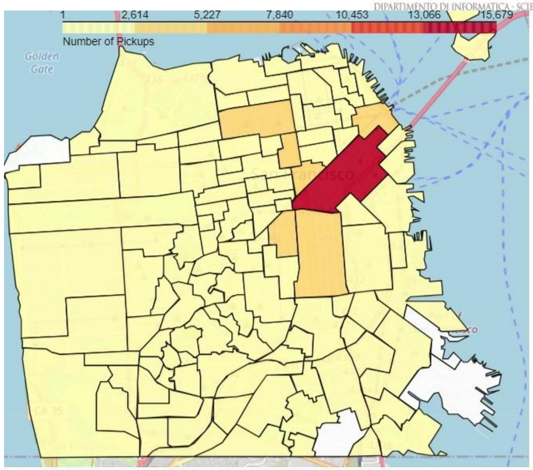
Generating choropleth maps: **ApproxGeoViz** Vs. baselines



reference choropleth map



ApproxGeoViz using **stratified-like** sampling



Baseline using **random** sampling (SRS)

sampled 8% → stringent stream settings → data arrival rates far exceeds system processing capacity

- For small sampling fractions, SRS **overlooks** more regions than stratified counterpart (colored in white in Figures)
- More accuracy is accrued by applying stratified-sampling in the front-stage

Outline

- ➢ Geospatial big data analytics: Background and Motivating scenario
  - o Motivating scenario
  - o Spatial data challenges & requirements
- ➢ A method for generating region-based approximate geo-maps
  - o Overview
  - o ApproxGeoViz
- ➢ Results and Discussion
  - o Deployment: baselines & testing setup
  - o ApproxGeoViz Vs. baseline
- ➢ Summary & future research

Concluding remarks

- **ApproxGeoViz** is a novel system for generation of approximate region-based geo-maps from voluminous georeferenced data
  - Stratified-like sampling quick-and-approximate filter to discards extra loads
- Employs information-theoretic EMD-based QoS **controller** to compute sampling rate
  - Loop **feed-back** mechanism
  - Guarantees sampling tuples that can be efficiently geo-visualized given the capacity of the system
- **Future research,** To develop a mathematically-principled algorithm to decide upon similarity value based on data stream statistics
  - Currently, arbitrarily-selected or expert-guided

جـامعــة الشـــارقــة
UNIVERSITY OF SHARJAH

Q&A and Contacts

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

Thanks for your attention!

Question's time...

Dr. Isam Mashhour Al Jawarneh[1],
Dr. Luca Foschini[2],
Prof. Antonio Corradi[2]

[1] *Assistant Professor*, Department of Computer Science, University of Sharjah, UAE (ijawarneh@sharjah.ac.ae)

[2] Associate Professor, Department of Computer Science and Engineering – DISI, University of Bologna, Italy (luca.foschini@unibo.it)

IEEE CAMAD 2023

06 - 08 November 2023 , Edinburgh, Scotland, Edinburgh Napier University