UNIVERSITÀ DELLA CALABRIA

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI INGEGNERIA INFORMATICA, MODELLISTICA, ELETTRONICA E SISTEMISTICA
DIMES

Methods, Tools and Applications for Scalable Data Analysis

NESUS Winter School - PhD Symposium 2017

Loris Belcastro

University of Calabria - DIMES

Research Outlines

Big Data

"High volume, high velocity, and/or high variety data that requires new forms of processing"

Cloud Computing

"Scalable storage and cost-effective processing services that can be used for extracting knowledge from Big Data repositories"

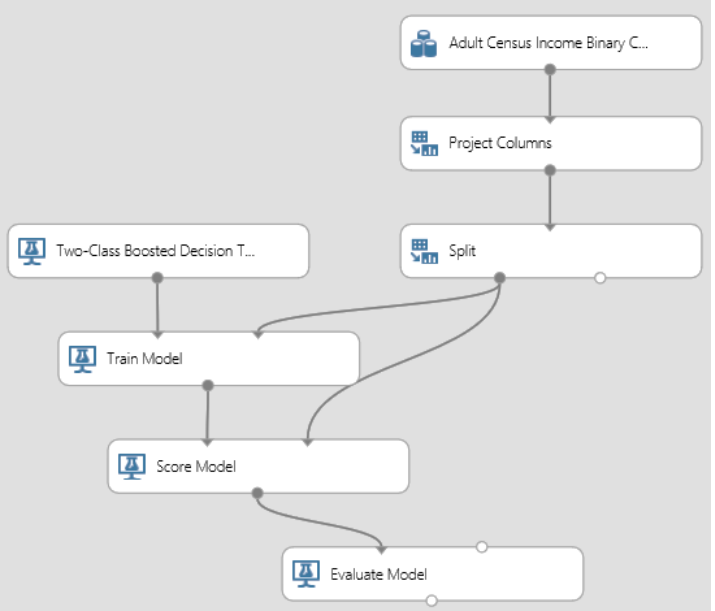
MapReduce

- *Programming model for writing parallel and distributed applications.*

Research Outlines

Data Analysis Workflows on Clouds

"A really effective way for expressing task coordination and creating data analysis applications."



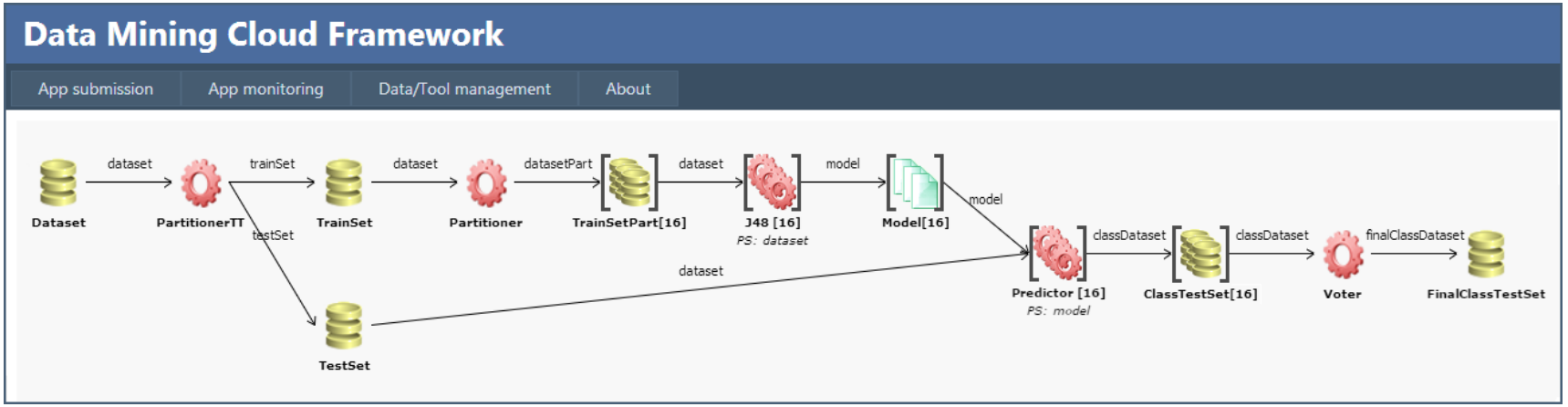
Social Data Analysis

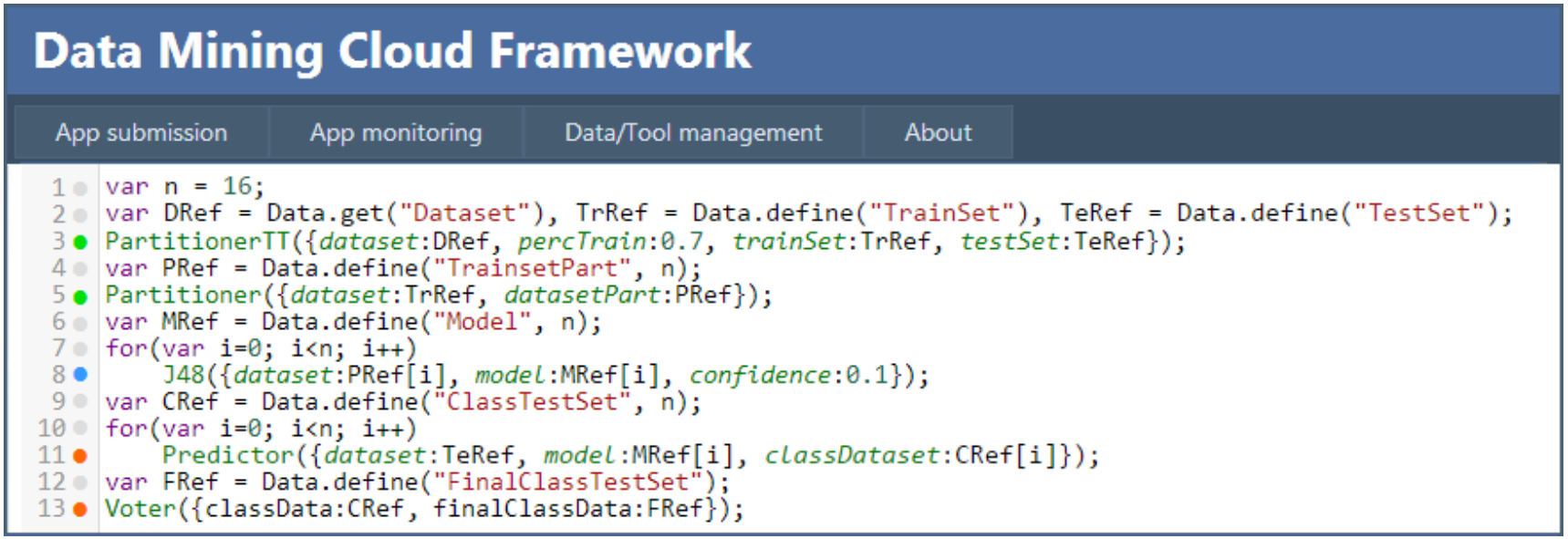
- *Extracting knwoledge from large datasets collected from social networks..*

Data Analytic Workflows

- The workflow paradigm is a compelling solution for representing complex data analysis applications
- Data Mining Cloud Framework (DMCF) is a platform, developed at the University of Calabria, for supporting the scalable execution of data analysis workflows on Clouds
- DMCF has been extended for integrating the MapReduce model into its workflow engine

DMCF – Visual- and script-based formalisms



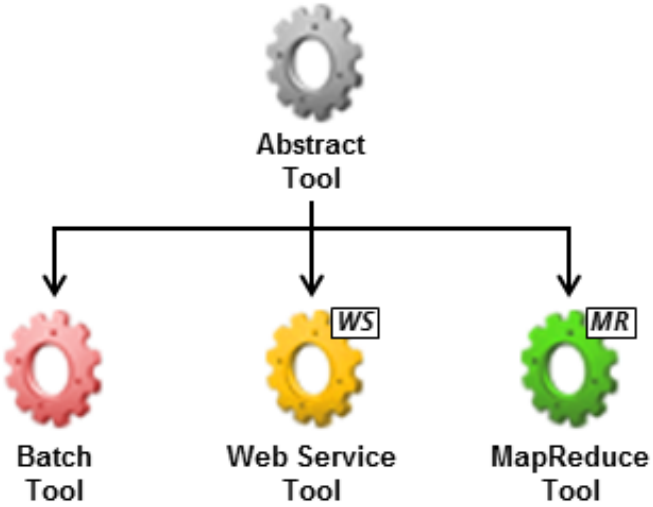


MapReduce in DMCF

- The use of the MapReduce model in workflows allows to improve the level of parallelism for some kinds of tasks, such as no iterative operations on large datasets.

- The concept of Tool (i.e., a Task) in DMCF has been generalized to support different types of tools

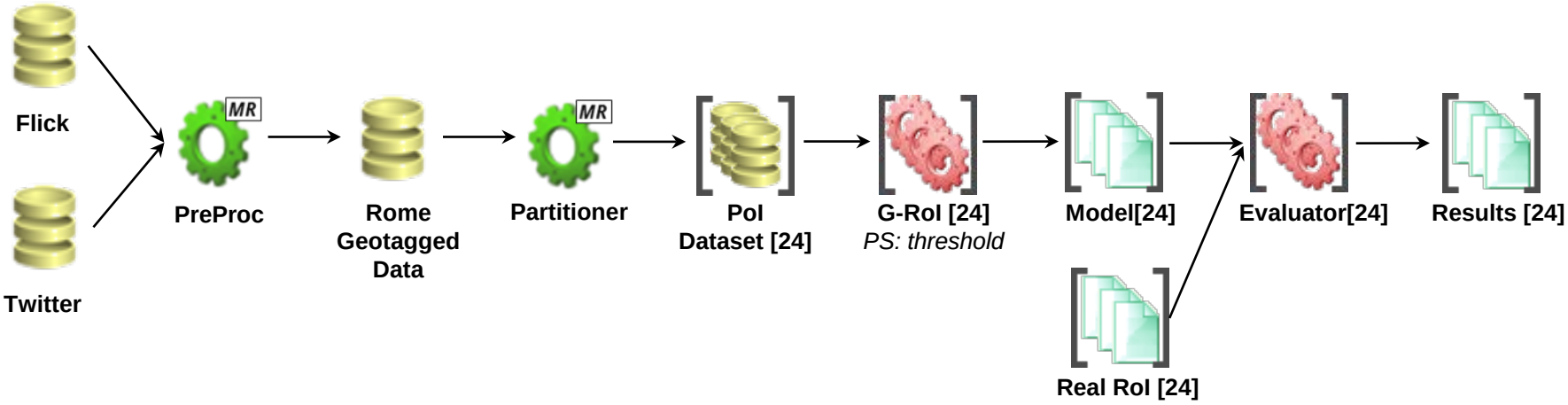
- Other frameworks for large data processing could be integrated in DMCF, expecially for executing iterative in-memory operations (e.g., Apache Spark).



MapReduce in DMCF – A study case

Extraction of Regions-of-Interest (RoIs) from geotagged social data:

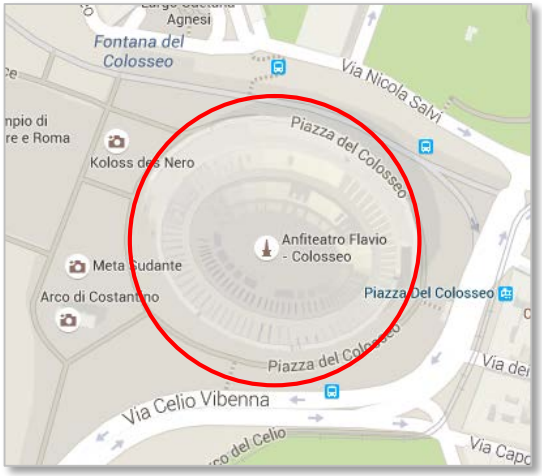
- About 500 GB of data from Twitter and Flickr
- Both MapReduce and batch tools have been used
- We used G-RoI, a novel RoI mining technique

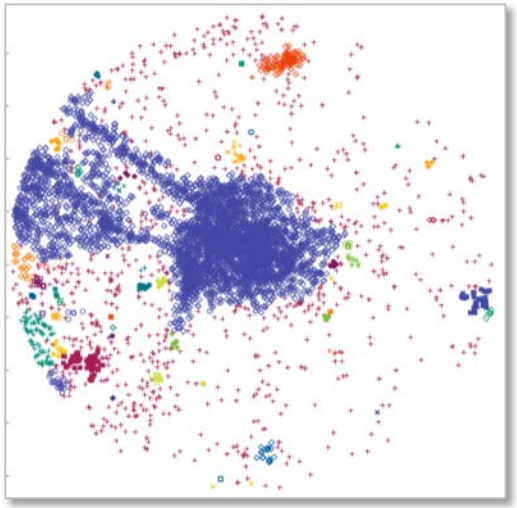


Existing approaches for Regions-of-Interest mining

Predefined shapes

This approach uses predefined shapes, such as circles of fixed radius, to represent RoIs.



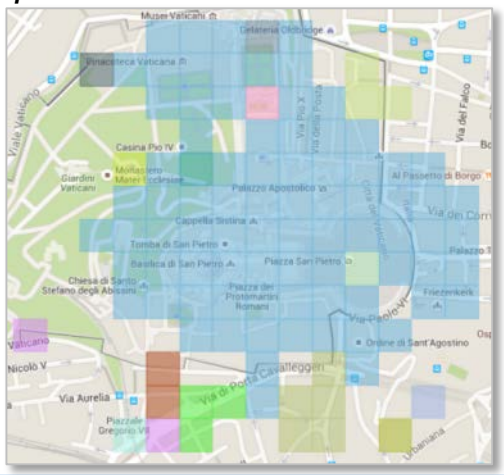


Density-based clustering

RoIs are obtained by exploiting density-based clustering algorithms, such as DB-SCAN or Optics, on a set of geographical points.

Grid-based aggregation

This approach discretizes the area under analysis in a regular grid and extract RoIs by aggregating the grid cells.



G-RoI – A novel tecnique for RoI mining

- G-RoI is a novel RoI mining technique that exploits the indications contained in geotagged social media items to discover RoIs with a high accuracy.


```
{ "id":"987654321",
  "owner":{"id":"123456789@N00","username":"FlickrUser"},
  "dateTaken":"May 3, 2015 4:39:24 PM",
  "tags":[
    {"value":"italy"},{"value":"rome"},{"value":"piazzadispagna"},
    {"value":"itali"},{"value":"spanishteps"}
  ],
  "title":"Night at Piazza di Spagna",
  "description": "In the Piazza di Spagna, just below the Spanish Steps",
  "geoData":{ "longitude":12.482045, "latitude":41.905888}
  ...
}
```


- It is easy to be configured
- Experiments performed over a set of PoIs in Rome and Paris demonstrated that G-RoI achieves better results than existing techniques.

G-RoI Methodology – Some definitions

- Let a PoI P be identified by one or more keywords $K=\{k_1, k_2, ...\}$.
  - As an example: the Colosseum is indeitified by the keywords “Colosseo, Coliseo, Colise, Coliseum, Amphitheatrum Flavium, Flavian Amphitheatre, etc”.
- Let G all be a set of geotagged items associated to P, i.e., the text or tags of each $g_i \in G$ contains at least one keyword in K.
- Let $C=\{c_1, c_2, ...\}$ be a set of coordinates, where c_i represents the coordinates of $g_i \in G$.
- Let cp_0 the convex-hull polygon that encloses all the coordinates contained in C.

G-RoI Methodology

Then, G-RoI follows two procedures:

G-RoI reduction:

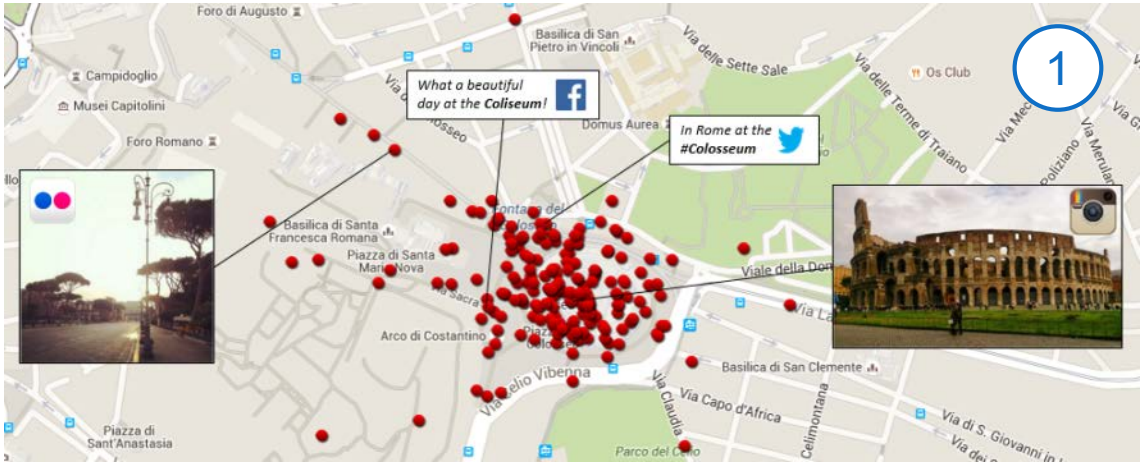
- starting from cp_0, it iteratively reduces the area of the current convex polygon by deleting one of its vertex using a density-based criterion.
- At each step, the procedure deletes the vertex that produces the polygon with highest density, among all the possible polygons.

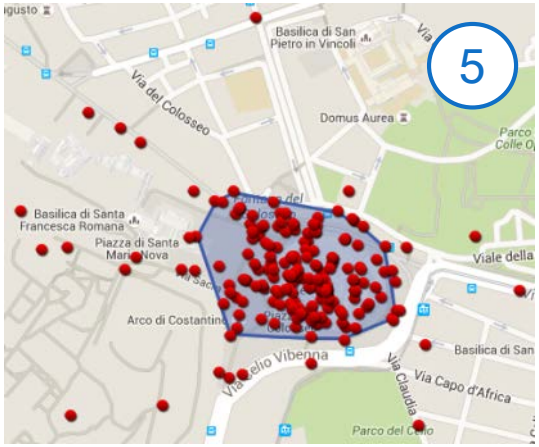
G-RoI selection:

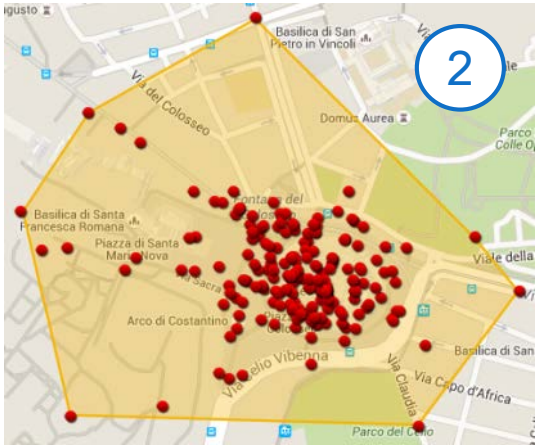
- It analyses the set of convex polygons returned by the reduction procedure, and selects the polygon representing the region of interest *R* for the PoI *P.*

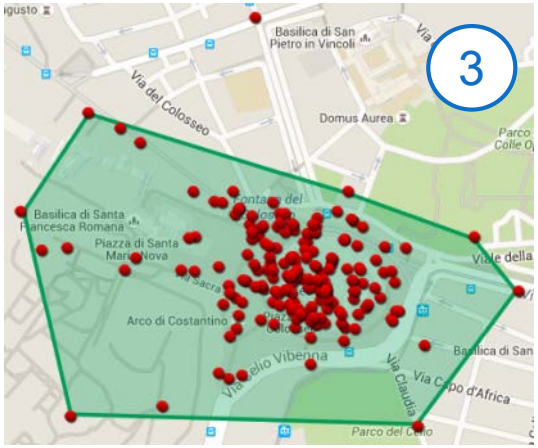
- An area-variation criterion is adopted to choose *R*. This corresponds to choosing cp_{cut} as the corner point of a discrete L-curve obtained by plotting the areas of all the convex polygons on a Cartesian plane.

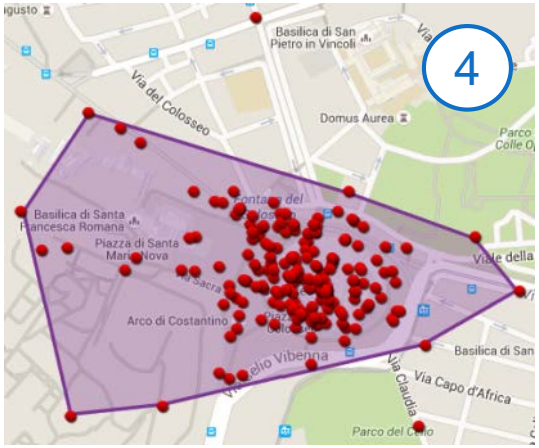
G-RoI Methodology – How does it works?





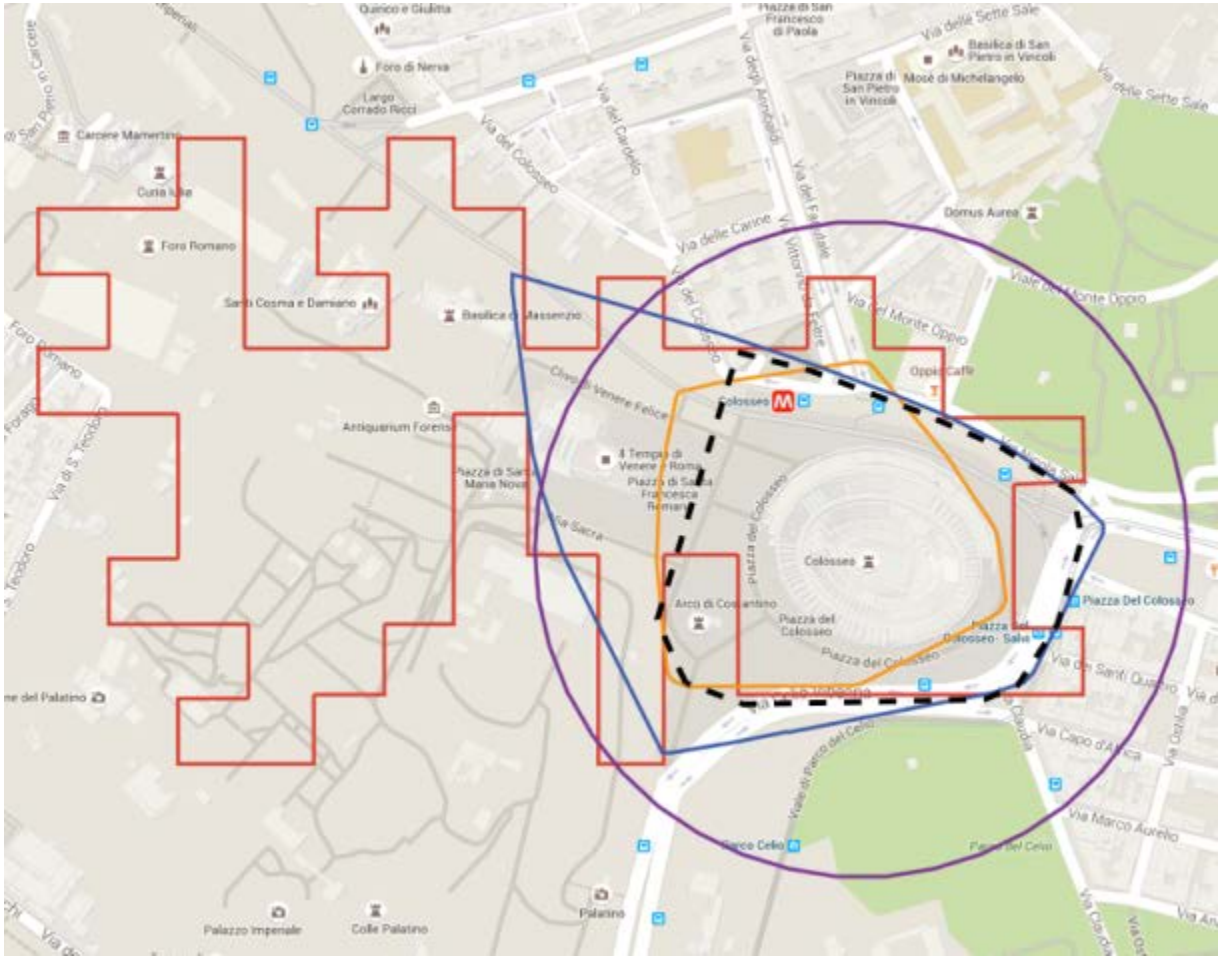






G-RoI – Comparison with other techniques

The Colosseum

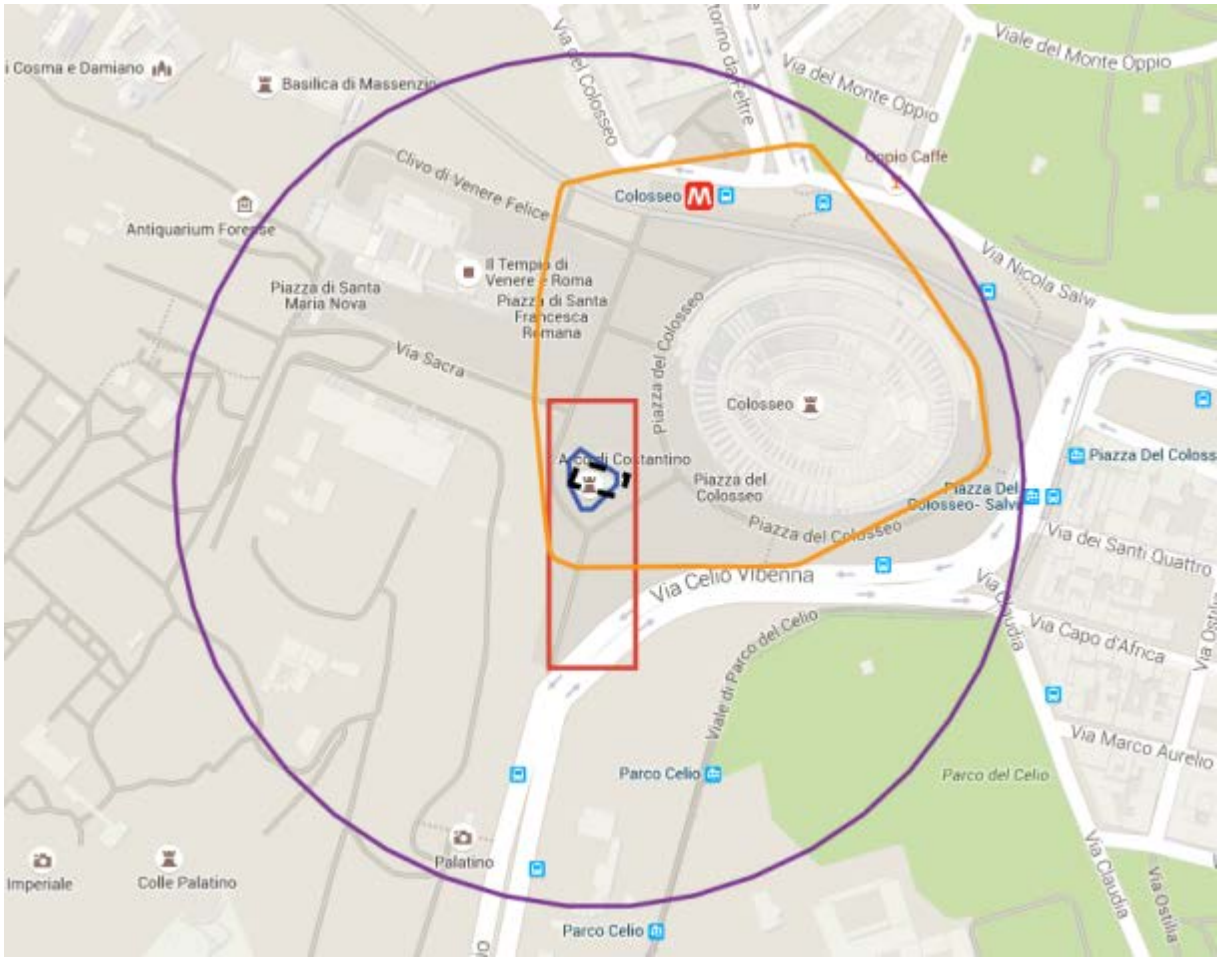


LEGEND

- **SlopeRoI**
- **Circle with a fixed radius**
- **DB-SCAN**
- **G-RoI**
- **Real RoI**

G-RoI – Comparison with other techniques

Arch of Constantine

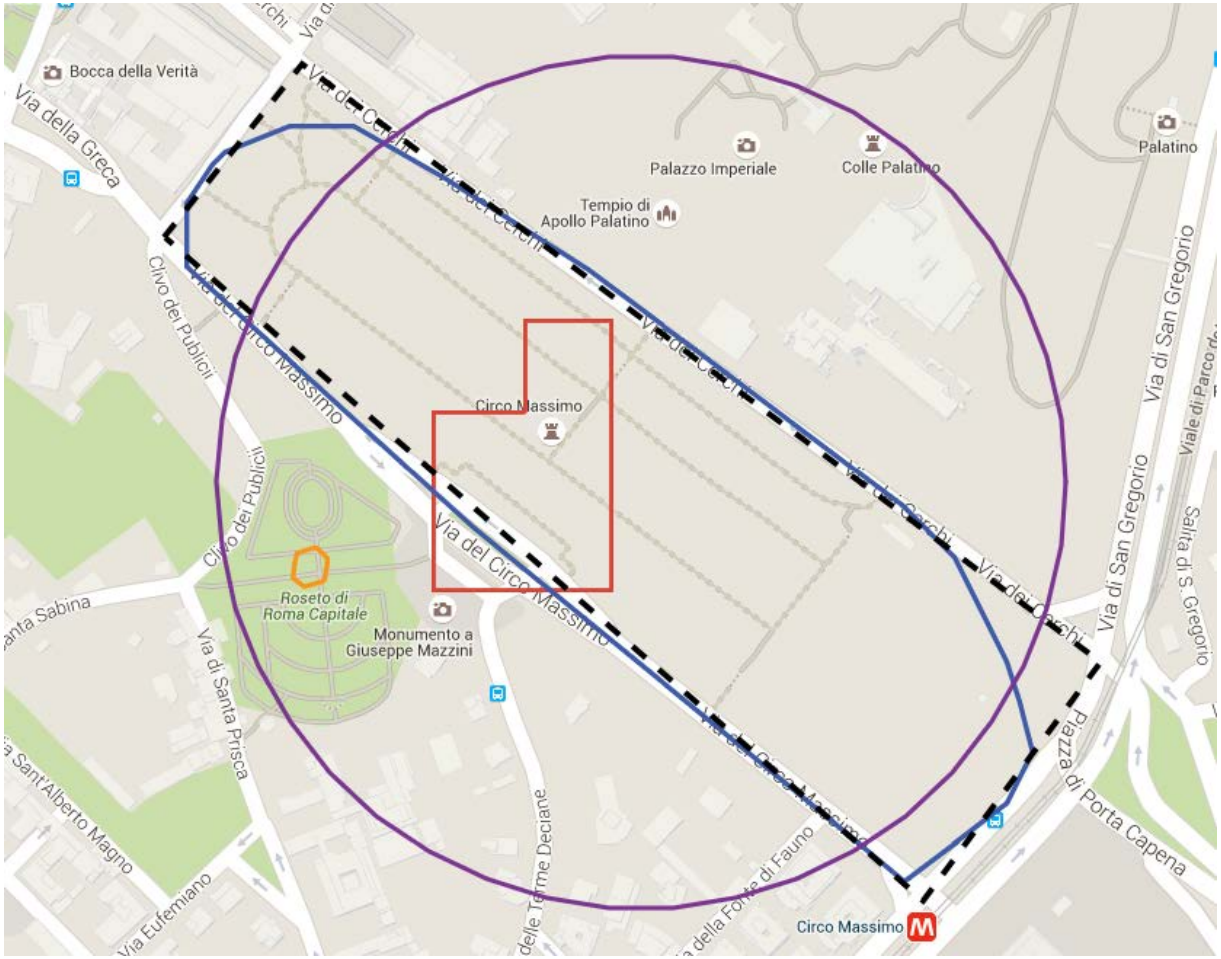


LEGEND

- **SlopeRoI**
- **Circle with a fixed radius**
- **DB-SCAN**
- **G-RoI**
- **Real RoI**

G-RoI – Comparison with other techniques

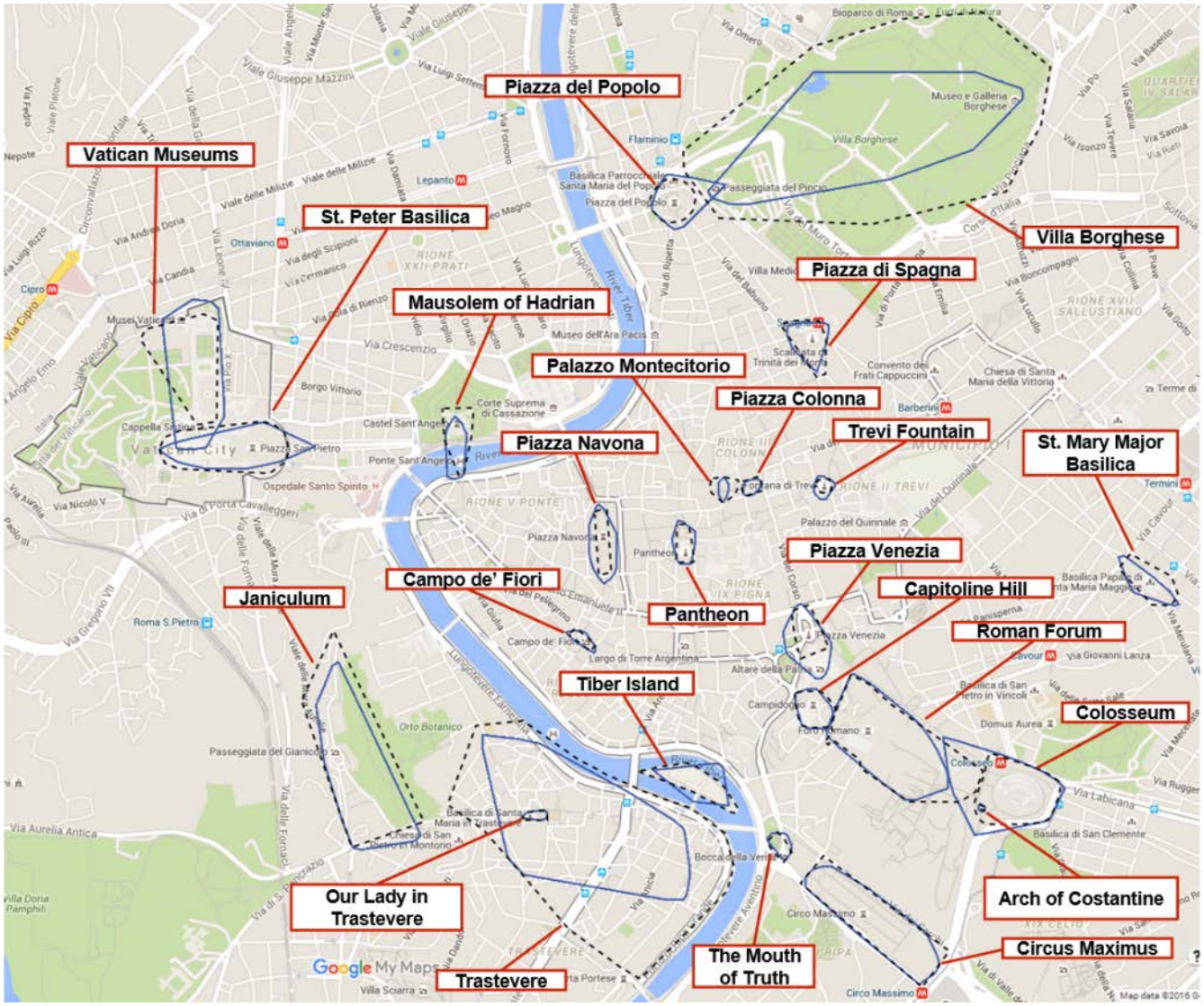
Circus Maximus



LEGEND

- **SlopeRoI**
- **Circle with a fixed radius**
- **DB-SCAN**
- **G-RoI**
- **Real RoI**

G-RoI – Results for 24 PoIs in the center of Rome



Thank you!

Loris Belcastro