Areas-of-Interest for OpenStreetMap with Big Spatial Data Analytics

SotM 2018 Milano - July 29th

sfkeller@hsr.ch

sfkeller@hsr.ch

HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL

FHO Fachhochschule Ostschweiz

Intro

Areas-of-Interest for OpenStreetMap with Big Spatial Data Analytics

- Areas-of-Interest (AOI) – State-of-the-Art
- AOI – Definition
- AOI with OSM: Implementation and processing steps
- AOI – Further work
- What about big spatial data?

Who knows what AOI are on Google Maps and how they look like?
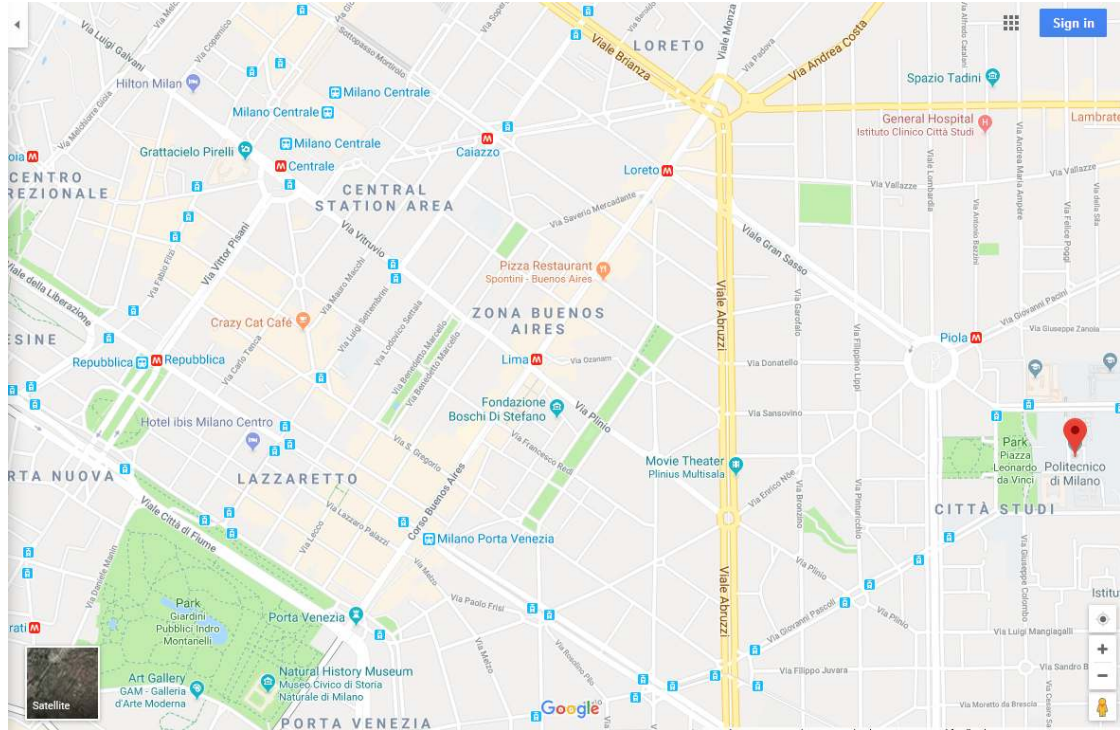
About Areas-of-Interest (AOI)

Many notions of AOI:

- "Computer-Assisted Editing": Areas with presumed missing data to be mapped in OSM, e.g. preselected areas for editing or specificly core areas for crisis mapping.

- „Tourism": shopping, entertainment and cultural areas to help travellers to explore the world.

- …

Let's take a glimpse where we are here in Milano in the quarter "Città Studi" plus "Buenos Aires-Venezia" westward!

AOI State-of-the-Art: Google Maps

Def. of AOI from a blog post:

● "places where there's a lot of activities"

● "areas with the highest concentration of restaurants, bars and shops."

● "In high-density areas like NYC, we use a human touch (…)." (July 2016, https://blog.google/products/maps/discover-action-around-you-with-updated/

● See shaded orange areas; single category: probably using user tracks

https://goo.gl/maps/ReFHjDWaoY82

4

AOI State-of-the-Art cont'd.: AVUXI.com

TopPlace™
Heat Maps Tiles

Based on OSM, Flickr, etc.

Barcelona based startup

Categories:

● Shopping (<< shown)

● Sightseeing

● Eating

● Nightlife

http://www.avuxi.com/heat-maps-demo

AOI State-of-the-Art cont'd.: AVUXI.com

TopPlace™
Heat Maps Vector

Categories:

- Shopping (<< shown)

- Sightseeing

- Eating

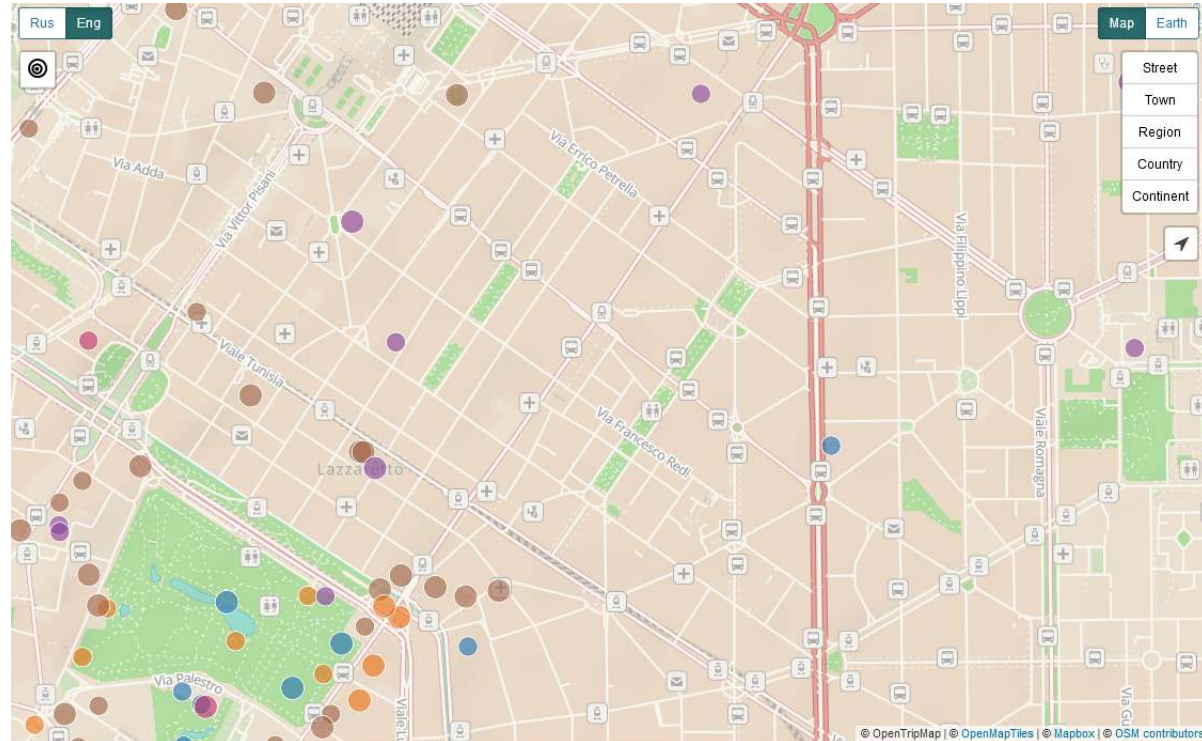- Nightlife

- Parks & Waterfront
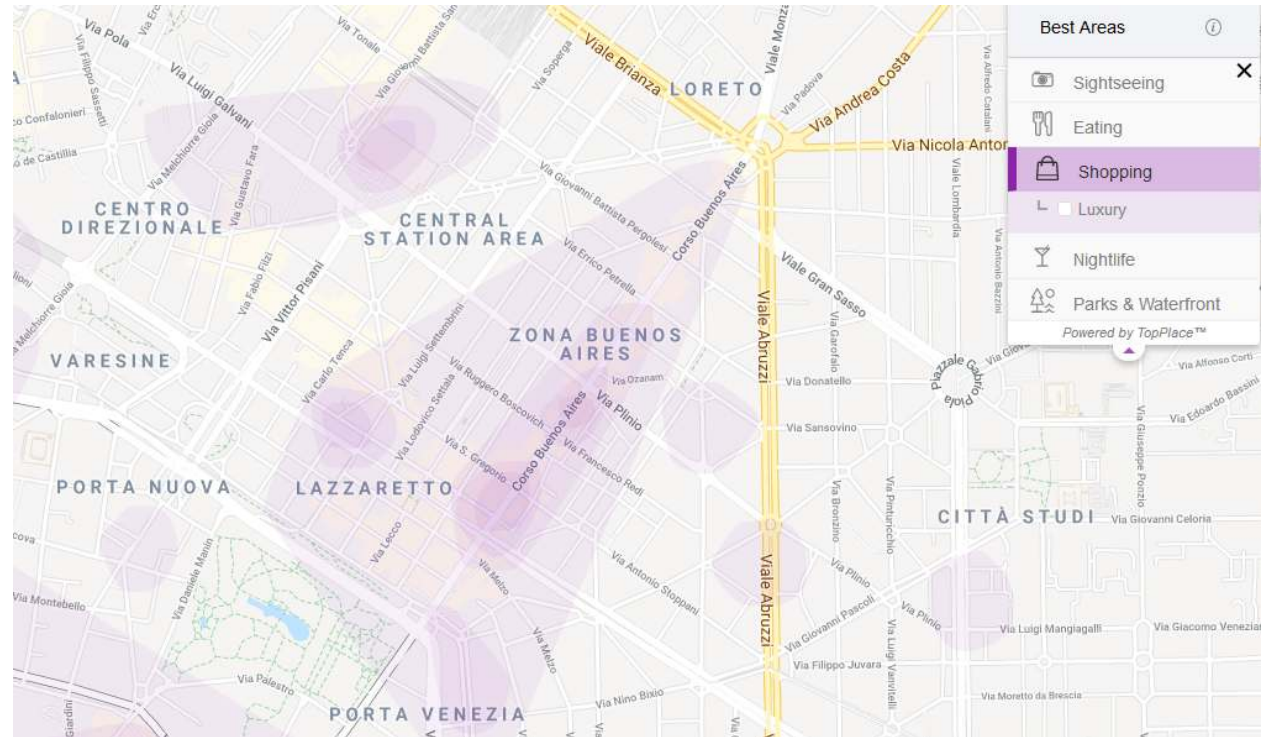
AOI State-of-the-Art cont'd.: OpenTripMap

Note: That's AOI at POI level (FYI!); it's not AOI at area level as we are interested in!

Based on OSM.

Criteria „Very famous"
Categories:
- Interesting Places
- Amusements
- Tourist facilities
- Accomodations

https://opentripmap.com/en/#15.5/45.4789/9.2112

Areas-of-Interest

- Our definition:
  "*Urban area at city or neighbourhood level with a high concentration of POI, and typically located along a street of high spatial importance*"

- Focus on neighborhood-level - not building level
- Focus on an aggregated category (includes sightseeing, eating, shopping, nightlife, leisure)

- Based on OpenStreetMap data
- and on a open documented, reproducible algorithm/process

AOI with OSM: Implementation

- Use Case as part of a Master Thesis by Philipp Koster, MSc Computer Science, HSR Rapperswil, Spring 2018 (see eprints.hsr.ch)

- Implement AOI and explore it's limits with
  - Open Source Software
  - PostgreSQL / PostGIS (spatial) SQL database
  - Python as data analytics programming language
  - other libraries / tools if needed

AOI with OSM: Processing Steps

1. Get polygons from OSM with/containing selected tags

2. Cluster polygons

3. Create hulls arround clusters

4. Apply network centrality using street network from OSM, extend hulls ~50m

5. Exclude water/waterways and sanitize

Done!

AOI Proc. Step 1/5: Get polygons from OSM

Get all polygons from OSM with/containing tags
Select polygons which:
• have a given tag
• contain a node with a given tag (and building = true)
• have not the attribute access = private
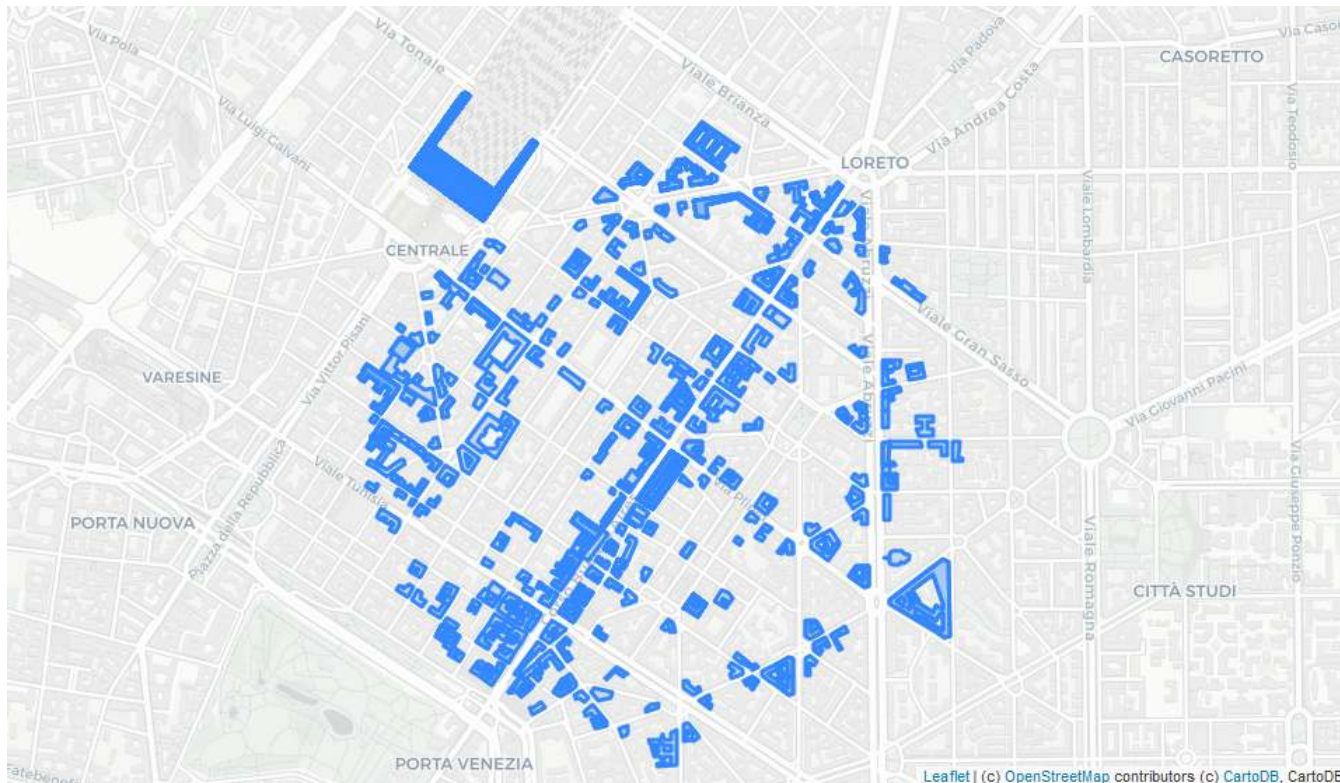
AOI Proc. Step 1/5: Get polygons … ff.

- Get polygons from OSM with selected tags

- Currently 87 tags

- See some of the selected tags here:

- landuse: retail
- amenity: cafe, restaurant, pharmacy, bank, fast_food, hospital, pharmacy, arts_centre, cinema, theatre, post_opffice, townhall, …
- shop: mall, bakery, healthfood, supermarket, boutique, jewelry, shoes, watches, hairdresser, ticket, laundry, tobacco, …
- leisure: amusement_arcade, beach_resort, fitness_centre, garden, ice_rink, sports_centre, water_park, …

12

AOI Proc. Step 2/5: Cluster polygons

Cluster polygons by using DBSCAN algorithm

DBSCAN parameters minPts and eps are locally adapted

ST_ClusterDBSCAN uses 2D impl. of "Density-Based Spatial Clustering of Applications with Noise"
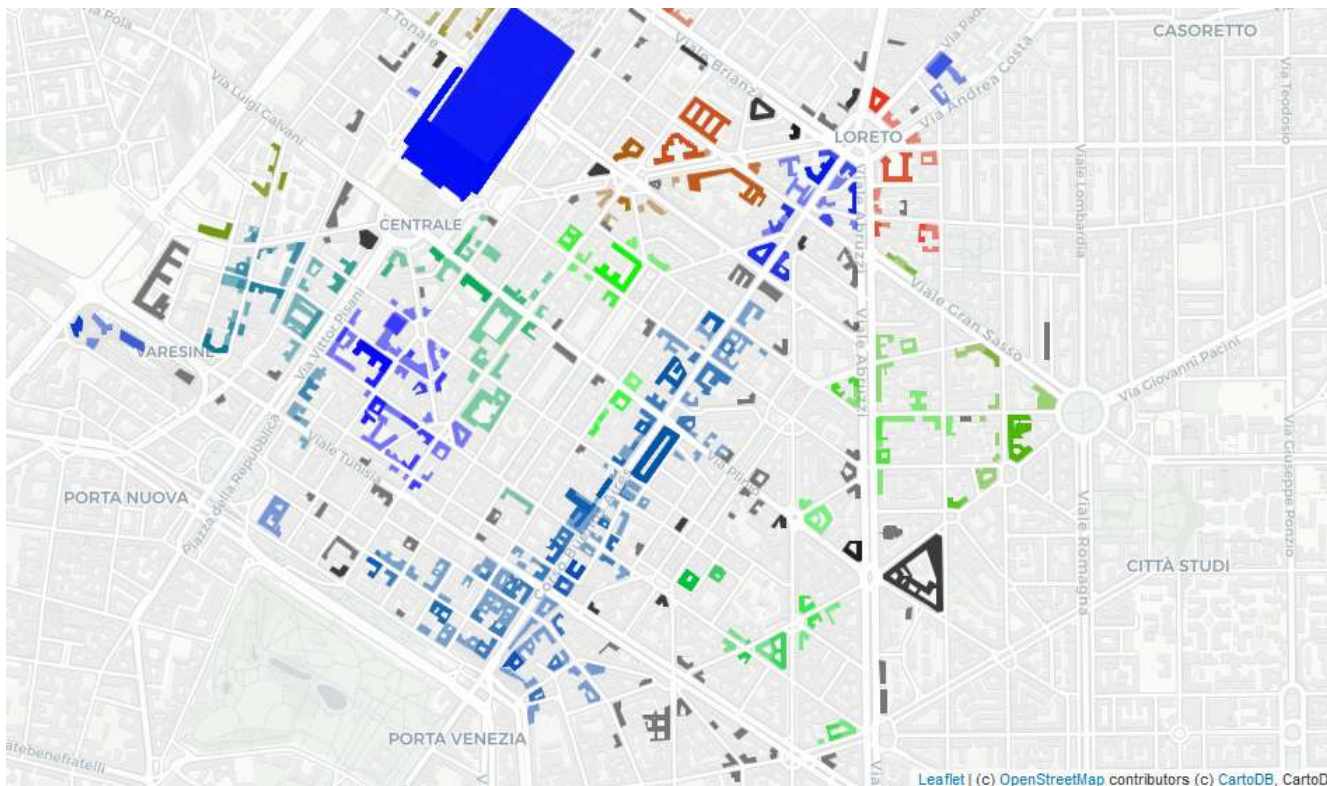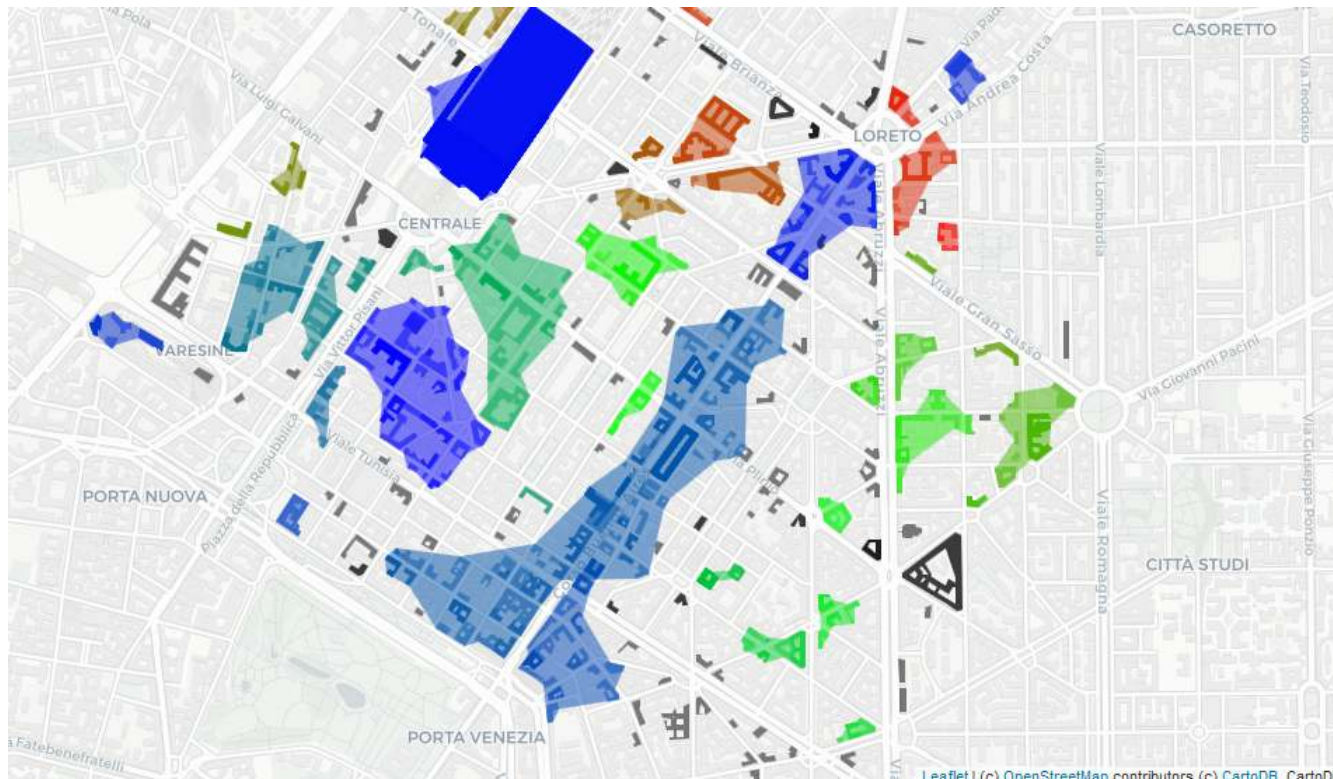
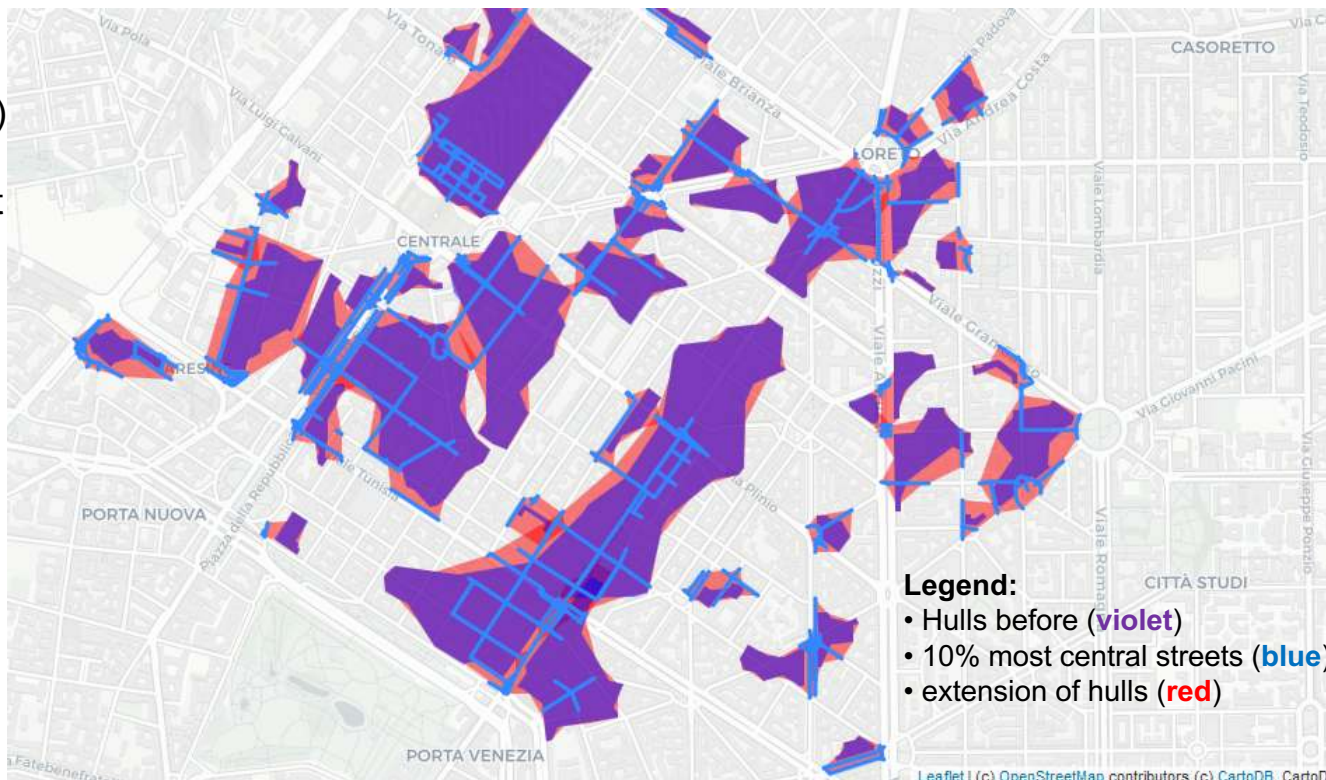# AOI Proc. Step 3/5: Create hulls around clusters

Concave hull

Using target_percent value of 0.99 (the target percent of area of convex hull)

Concave preferred over convex hulls

AOI Proc. Step 4/5: Apply network centrality

• Calculate closeness centrality of street graph for each hull (incl. buffer)

• Select 10% of the most central streets and ways

• Cut streets which are leaving the hull after 50 meters

• Extend hulls by drawing concave hull arround hull and (selected and cut) streets

Legend:
• Hulls before (**violet**)
• 10% most central streets (**blue**)
• extension of hulls (**red**)

15

AOI Proc. Step 5/5: <u>Exclude water</u> & sanitize

if water/waterways are present!

(not the case in Milano between quarters "Città Studi" and "Buenos Aires-Venezia")

AOI Proc. Step 5/5: <u>Exclude water</u> & sanitize
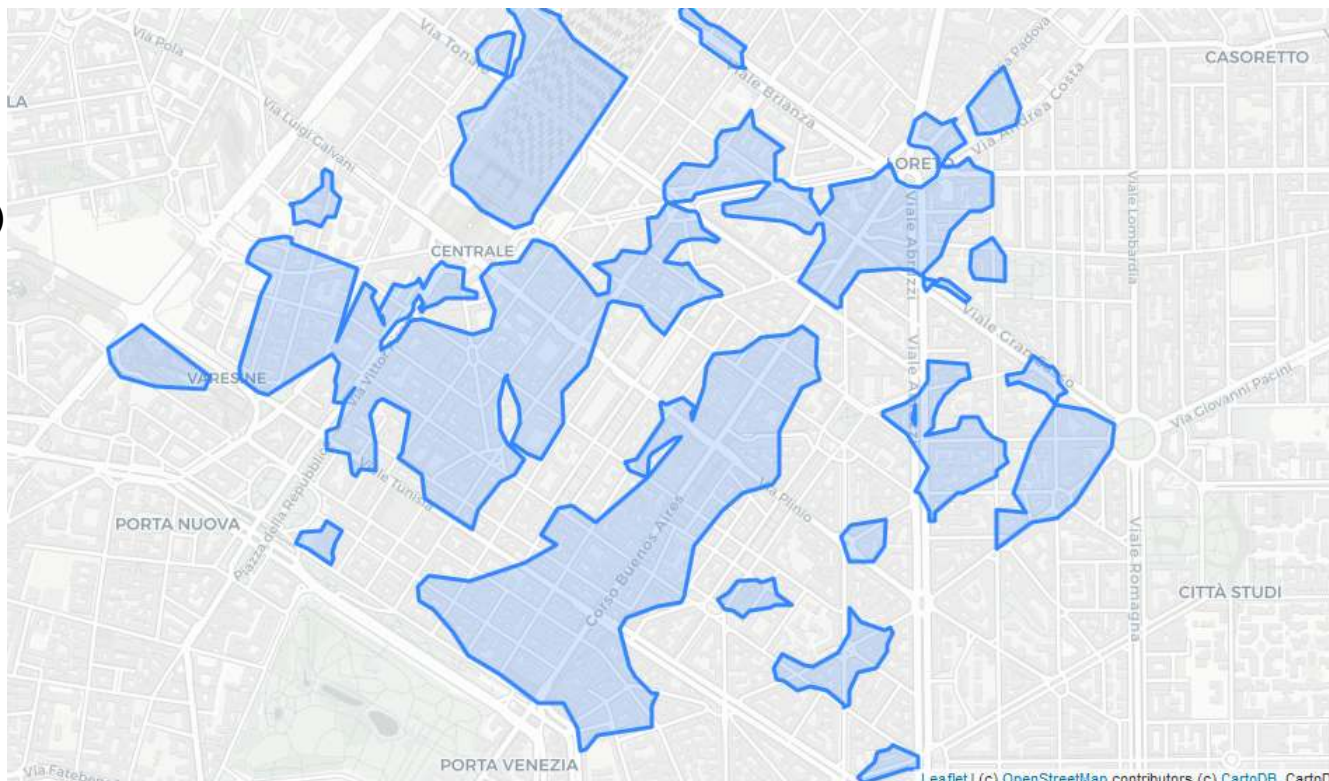
if water/waterways are present!
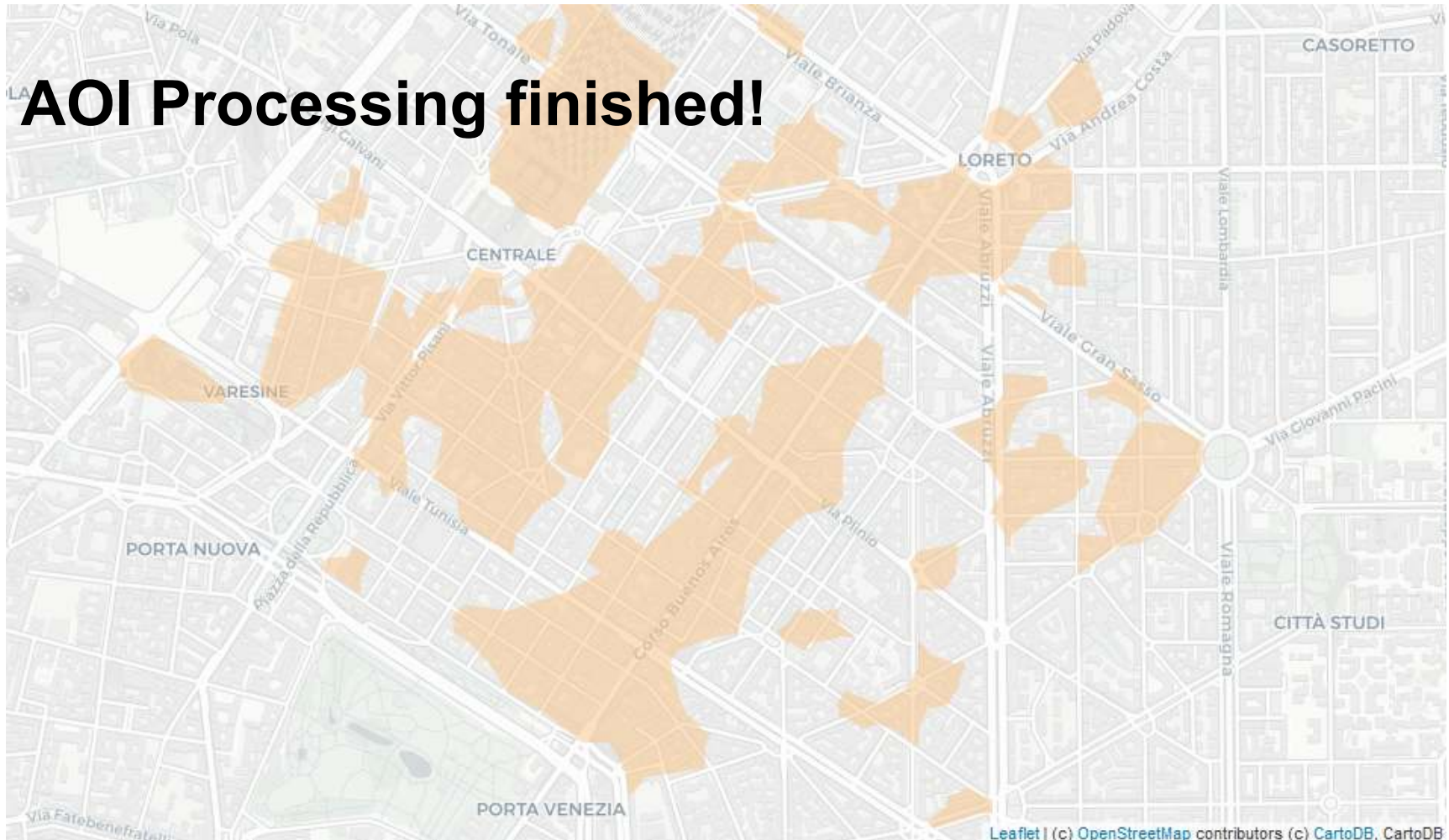
In Zürich old town there's water…
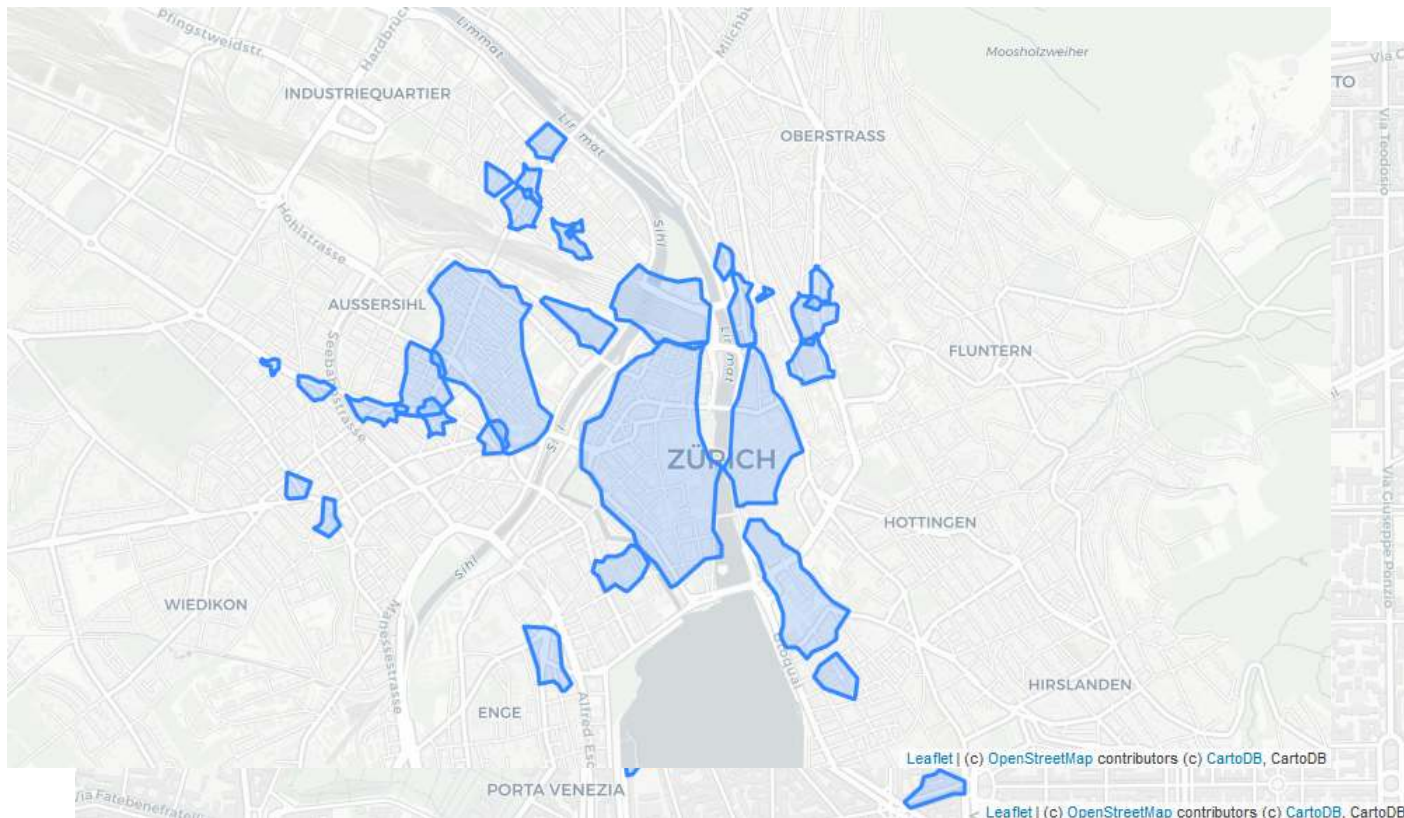
AOI Proc. Step 5/5: Exclude water <u>& sanitize</u>

Sanitize:
- Union overlapping polygons (ST_Union)
- Simplify polygons slightly (ST_Simplify(5))
- Remove invalid polygons (ST_IsValid and not ST_IsEmpty)

AOI Processing finished!

19

Evaluation - Discussion

Success! Justin O'Beirne essay 2017: "Google Maps's Moat - How far ahead of Apple Maps is Google Maps?"
https://www.justinobeirne.com/google-maps-moat : "It's no longer enough to simply collect data. Now to compete with Google, you also have to process that data (…). It's also interesting to ponder what this means for OpenStreetMap."
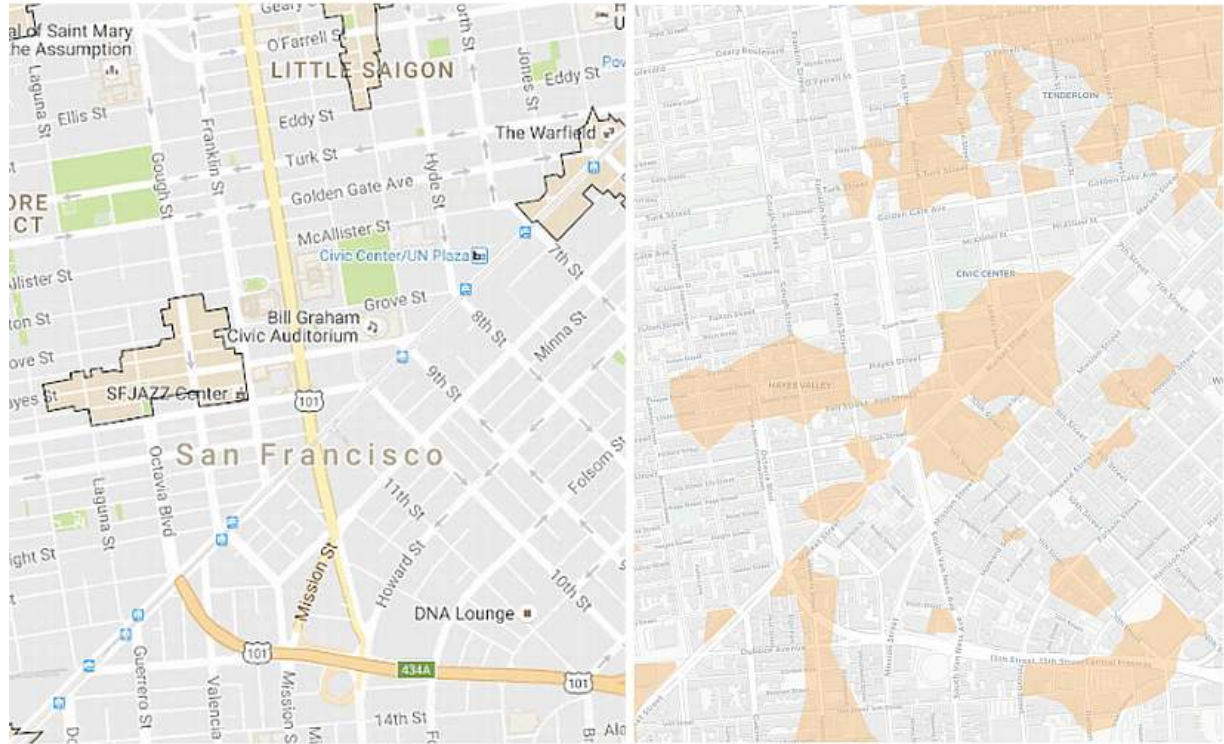
Further work on AOI - Discussion ff.

- Theses of SK52
  - "Can we identify 'completeness' of OpenStreetMap features from the data?" by SK53, 24 July 2018, http://sk53-osm.blogspot.com/2018/07/can-we-identify-completeness-of.html
  - AOI can be generated for less well off parts of town
  - Parametrisation means that even incomplete mapping can help
- Optimize local adaption of DBSCAN parameters
- More input data?

Technologies

- Python, the computer language
- PostGIS (PostgreSQL), open source database
- OSMnx, Python open source library for street networks based on OSM
- Jupyter Notebook, publishing format and interactive environment for reproducible computational workflows
- Docker, containerization software

Web resources

- AOI demo web page:

  - on demand (mail me ☺)

- AOI open source:

  - on github https://github.com/geometalab/ (soon)

- Master thesis (including AOI):

  - on university repository https://eprints.hsr.ch > Philip Koster

- AOI data of Switzerland (as GeoJSON):

  - on open research data publishing platformas DOI
    https://doi.pangaea.de/10.1594/PANGAEA.892644

What about Big Spatial Data?

- Other use case of Master Thesis by Philipp Koster

- Implement AOI with OSM using a "Big Data Framework" with

  - Open Source

  - SQL if possible

  - and with other libraries / tools if needed

- Spark-related project candidates which focus on SQL and vector data:

	GeoWave	GeoMesa	GeoTrellis	GeoSpark	STARK	Magellan	LocationSpark	GeoPandas
Compatibility with Apache Spark	✔	✔	✔	✔	✔	✔	✔	✔
Spatial Analysis Methods with Vector Data	(✔)	✔	(✔)	✔	✔	✔	✔	✔
Support for SparkSQL	(✔)	(✔)	(✔)	✔	✗	(✔)	✗	✗
Spatial Partitioning	✔	✔	✔	(✔)	✔	✔	✗	✗
Efficient Spatial Joins	✔	✔	✔	✔	✔	✔	✔	✗
Comprehensive Documentation	(✔)	✔	✔	(✔)	✗	✗	✗	(✔)
Compatible with pySpark	✗	✗	✔	✗	✗	✗	✗	✔

AOI processing with Big Spatial Data

- Technologies chosen
  - GeoSpark
  - DataFrames (SQL+Scala)
  - with fallback to RDD (Scala)
- GeoSpark:
  - + Good documentation
  - + Efficient Spatial Joins
  - - No Support for PySpark
- Runner-up GeoMesa:
  - - Not completely designed with Apache Spark (though possible)
  - - More dependencies than GeoSpark (like e.g. Accumulo)
  - + Now probably larger community and higher activity

Lessons learned RDBMS vs. Apache Spark

- The RDBMS approach:
    - PostgreSQL und PostGIS are rock-solid implementations
    - Network Centrality is bottleneck being externals lib
    - Developping in SQL is a time-saver
- „The Apache Spark approach":
    - + Apache Spark: mature; comfortable tools
    - - Apache Spark: steep learning curve; many dependencies
    - - GeoSpark is buggy and lacks functionality (currently 8 „ST_"-functions)
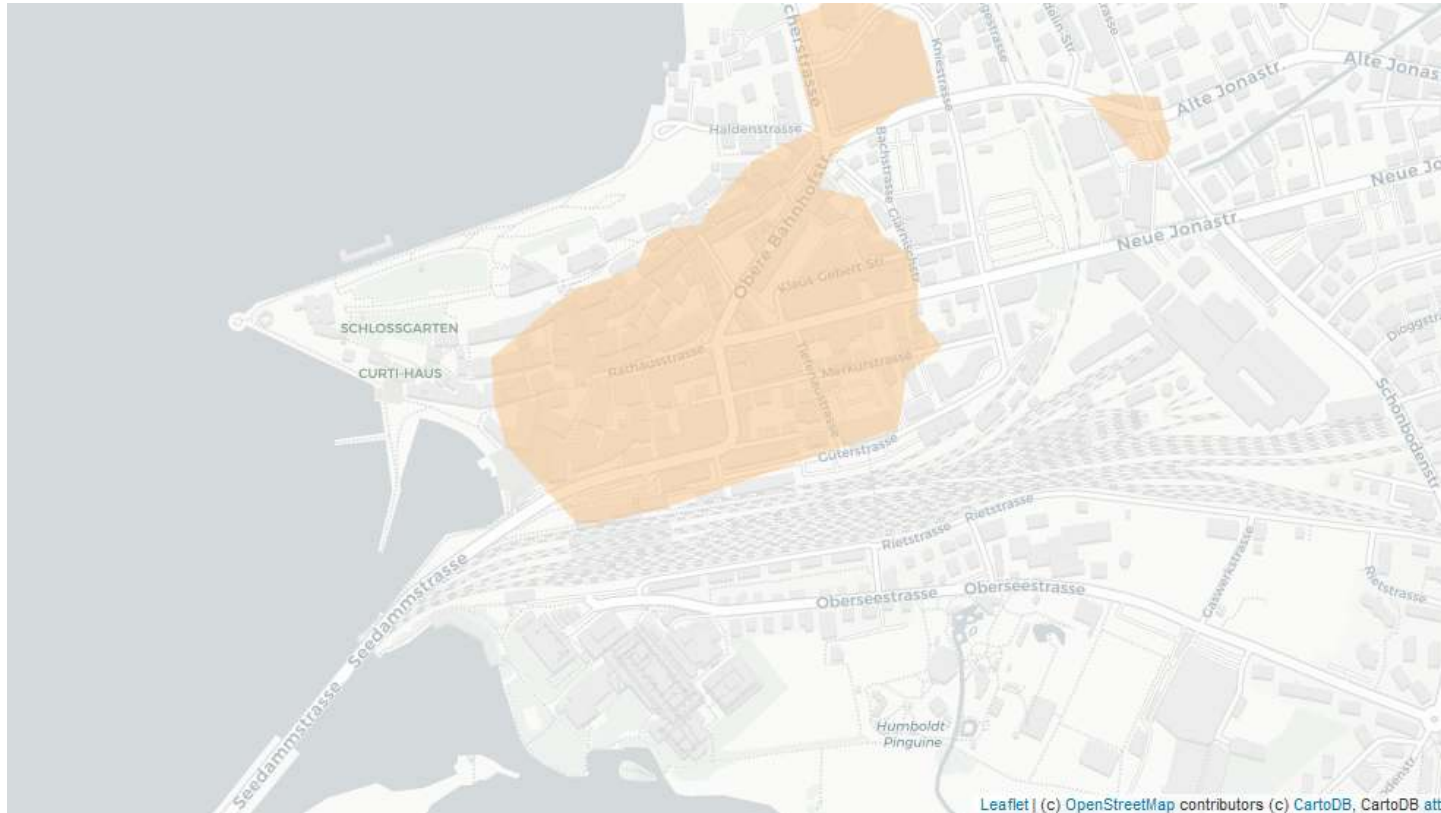   - - No performance gain (with data below 500 MB)

Thanks

- Philip Koster – master thesis https://eprints.hsr.ch and data (GeoJSON) https://doi.pangaea.de/10.1594/PANGAEA.892644 => my (former) student

- HSR – www.hsr.ch/geometalab => my Geometa Lab team at HSR

- Kang Zi Jing, Computer Science NTU, Singapore => former lab intern

- Jerry Clough, UK - http://sk53-osm.blogspot.com => active mapper

Questions?

AOI: Demo

Rapperswil
(Switzerland)