



# Extracting actionable information from unconventional data sources: social media analysis with VisualCit

Carlo Bono, Politecnico di Milano - carlo.bono@polimi.it
Barbara Pernici, Politecnico di Milano - barbara.pernici@polimi.it

May 3, 2022

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 872944*

Crowd4SDG

# Who am I

PhD candidate @ Politecnico di Milano (Department of Electronics, Information and Bioengineering)
My (super!) supervisor is Barbara Pernici

Background: Philosophy and Computer Science

Main research interests: data-driven applications, adaptive information systems, machine learning

Currently working on EU H2020 project Crowd4SDG
For support with the tools and/or movie recommendations: carlo.bono@polimi.it

# Why social media content analysis?

Social media contents are about «everything that is happening right now». So convenient when you need **information**!

But they're **huge**, and they're **noisy**. We want to retain **relevant** data only (in our dreams).

So we need **scalable** tools and algorithms that **understand** the contents of social media posts, in order to **filter** and **augment** them

Crowd4SDG

# Why social media content analysis?

Most of the functionalities that we are going to discuss are aimed at **image** data (hence the «Visual» in VisualCit), but many of the concepts are general

Often one (implicit?) goal will be to **reduce** the number of elements to be manually processed, while keeping **quality** high

# Crowd4SDG

Crowd4SDG

Crowd4SDG

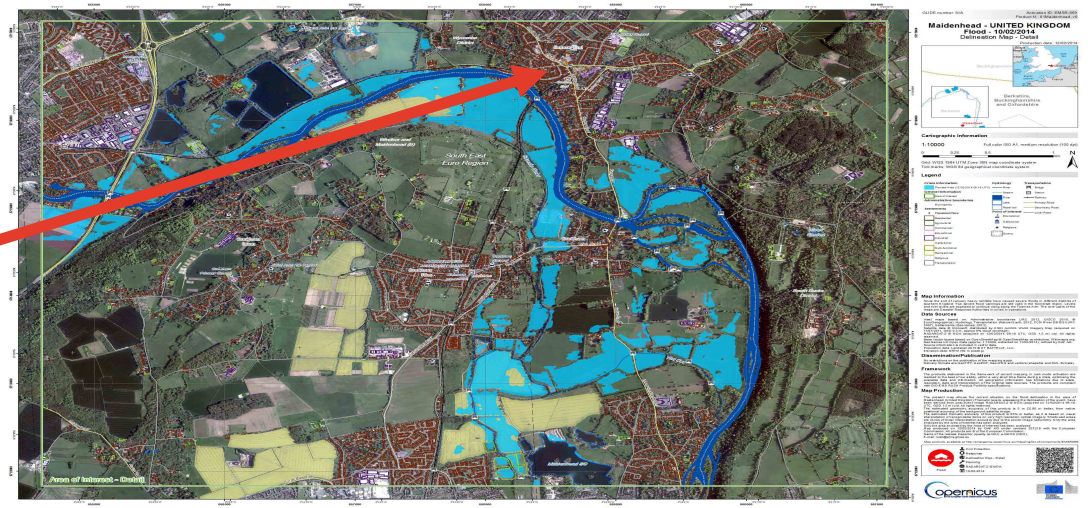
- https://crowd4sdg.eu/
- https://pernici.faculty.polimi.it/crowd4sdgpolimi/


- V. Negri, D. Scuratti, S. Agresti, D. Rooein, G. Scalia, J.L. Fernandez-Marquez, A. Ravi Shankar, M. Carman and B.Pernici, *Image-based Social Sensing: Combining AI and the Crowd to Mine Policy-Adherence Indicators from Twitter*, accepted at ICSE, Track Software Engineering in Society, May 2021 link
- Barbara Pernici, *CROWD4SDG: Crowdsourcing for sustainable developments goals*, 248-253, in Book of Short Papers, SIS 2020, Pearson, 2020 (link)

# Getting evidence from tweets

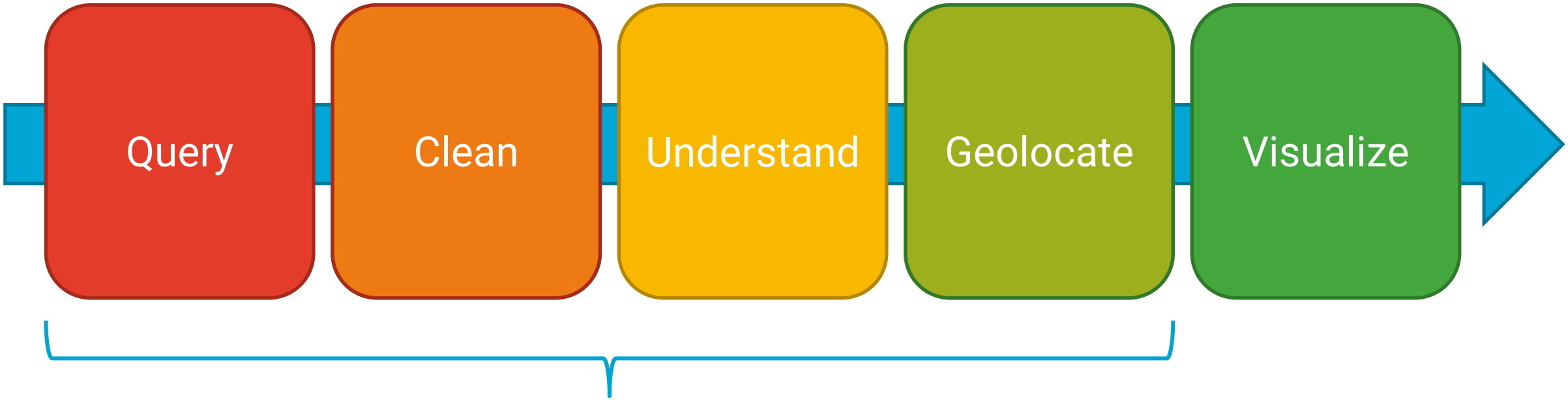
In a nutshell:
ingest social media, filter/enrich data, aggregate





Crowd4SDG

# A data analysis pipeline in practice




In some sense they all are ways of filtering

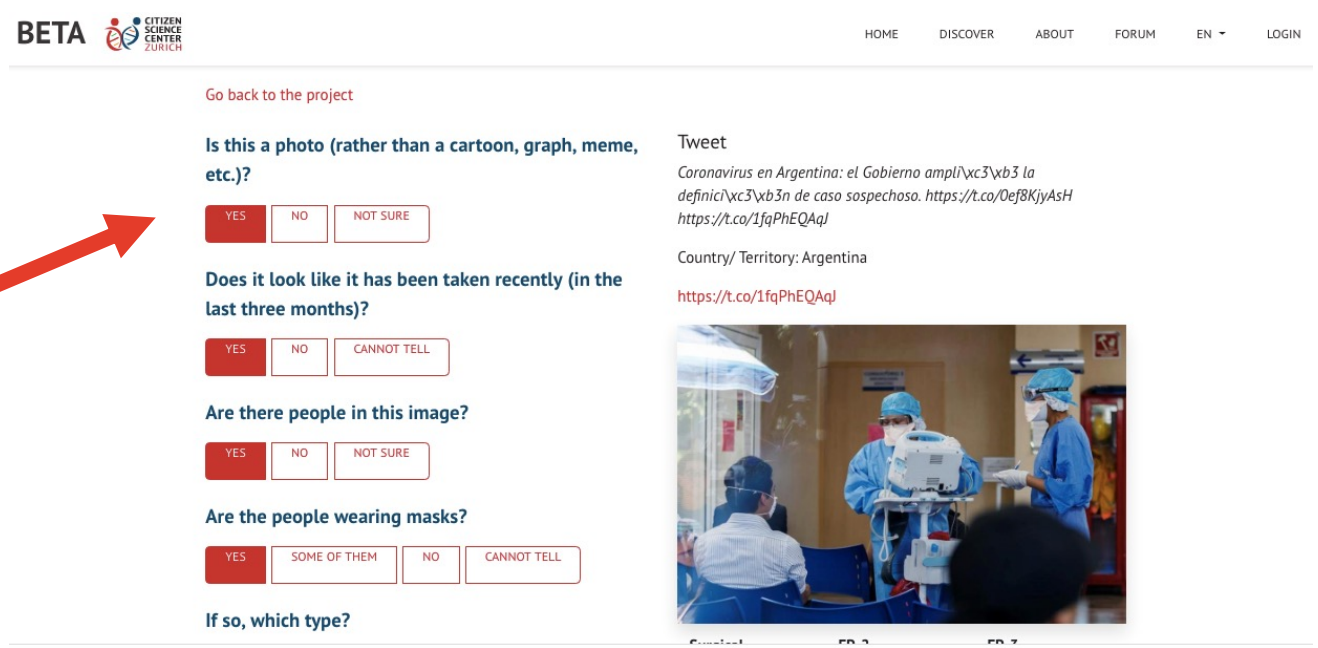
# «Helping machines do their work»

Crowd4SDG

PIPELINE

- Twitter keyword Crawl
- Duplicate URL Filter
- Similar Image Filter
- Photo Detector
- NSFW Detector
- Scene Detector
- Person Detector
- Geocoding
- Crowdsourcing
- Aggregate Statistics
- Visualisation on Map

## CROWDSOURCING



# VisualCit pipeline
## step by step

Crowd4SDG

- Which keywords?
- Do keywords change over time?
- How much data?
- Usually done via an API

# Crawling Social Media (e.g. by keywords)

Crowd4SDG

- Don’t do the same work again and again!
- Different meaning for image and text

Duplicated URL
Identical / Similar Image Filter

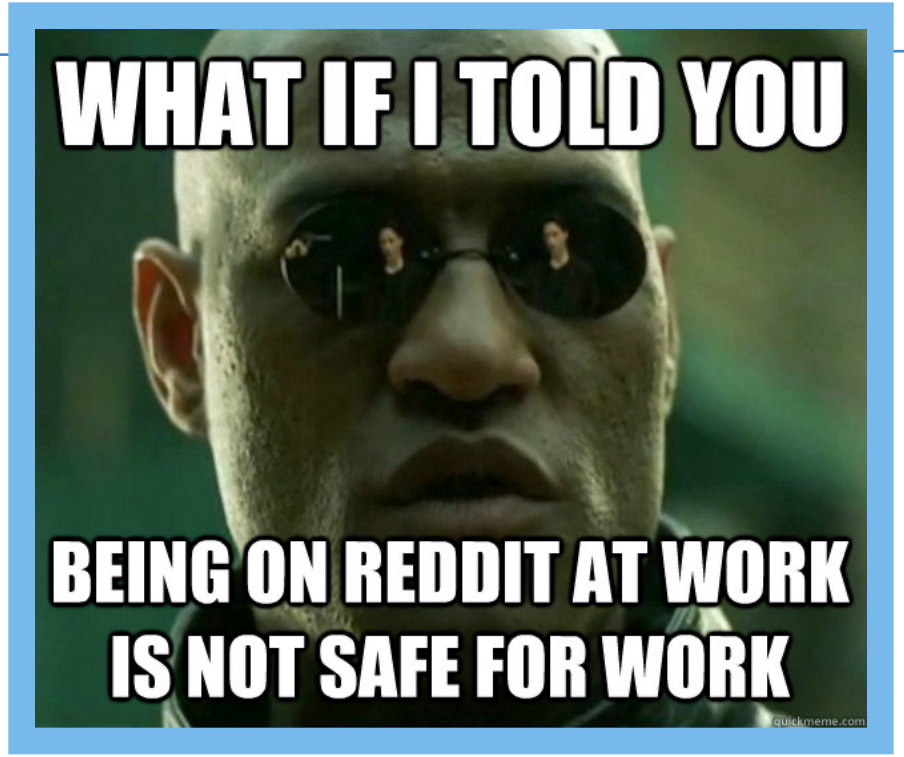
Crowd4SDG

- Drawings
- Documents
- Memes
- Graphs
- Screenshots


Remove “noisy” images
(e.g. non–photos)

Crowd4SDG

- Usually not of interest




Remove NSFW images

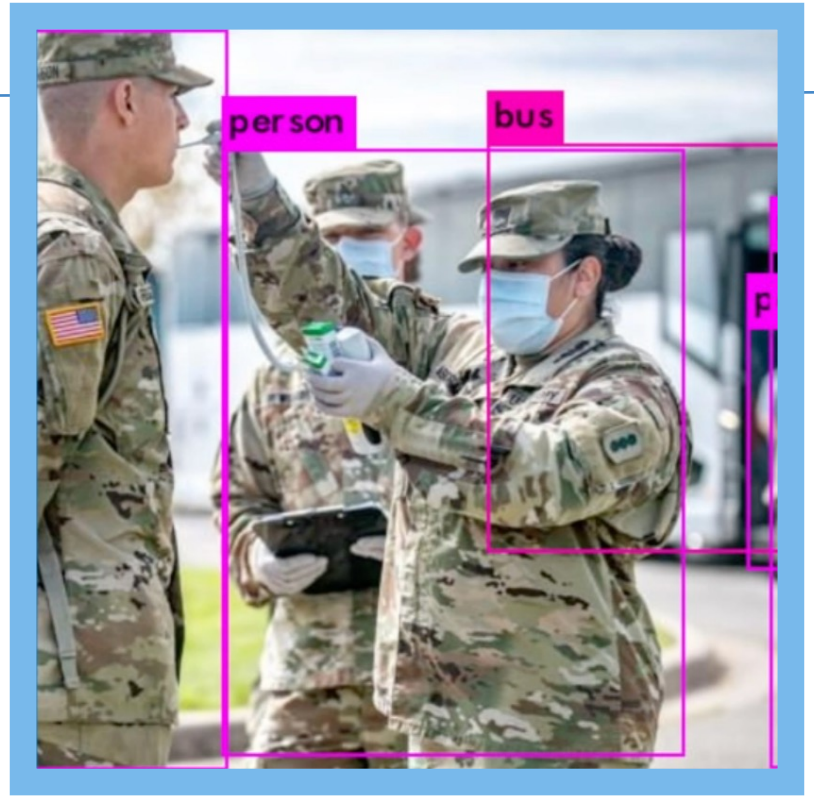
Crowd4SDG

- Some off-the-shelf classifiers, like scene classifiers
- Or can be custom


Select images of interest by kind: scene classification

Crowd4SDG

- Again, there are off-the-shelf object detectors
- Or it be custom, but it can take some effort




# Select images of interest by content: object detection

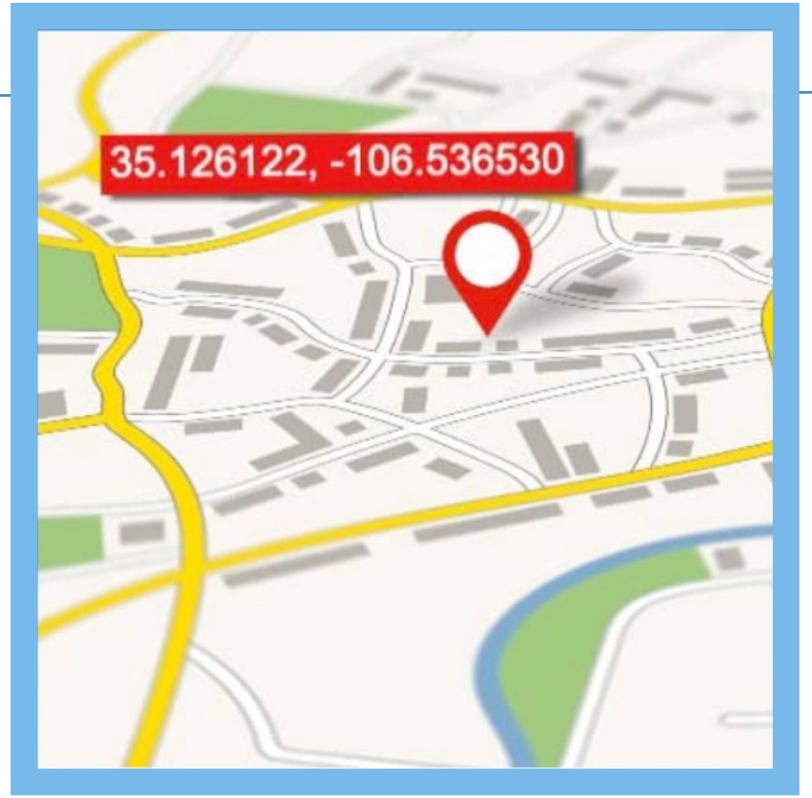
- Yep, it’s water!

Source: floodfinder.org


Analyse image regions by content: image segmentation

Crowd4SDG

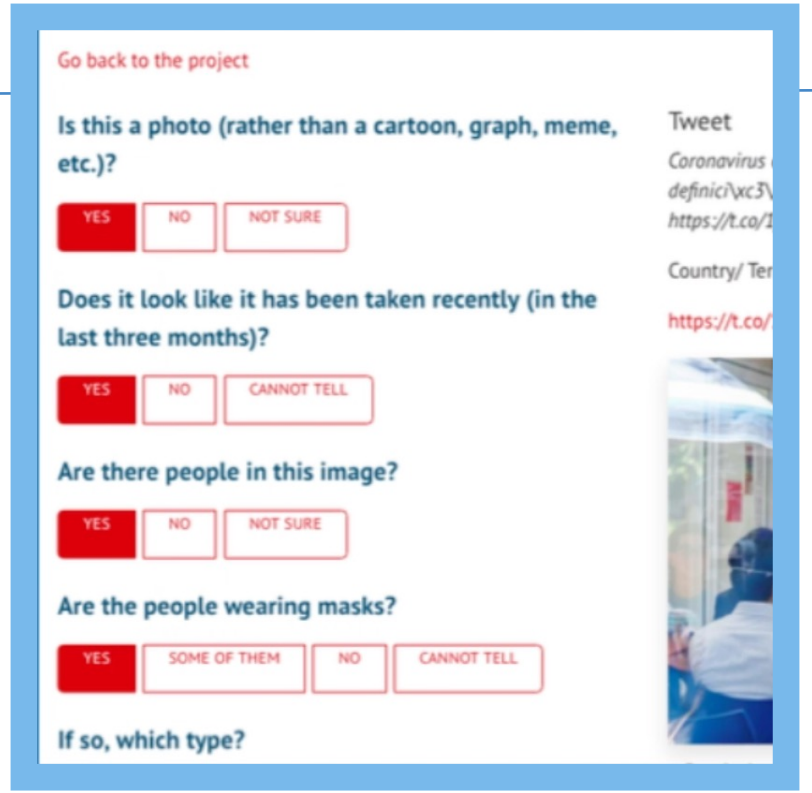
- Use some features of the posts and try to infer the location (not trivial!)




- Different applications have different constraints (e.g. granularity)

# Geocoding

Crowd4SDG




# Crowdsourcing

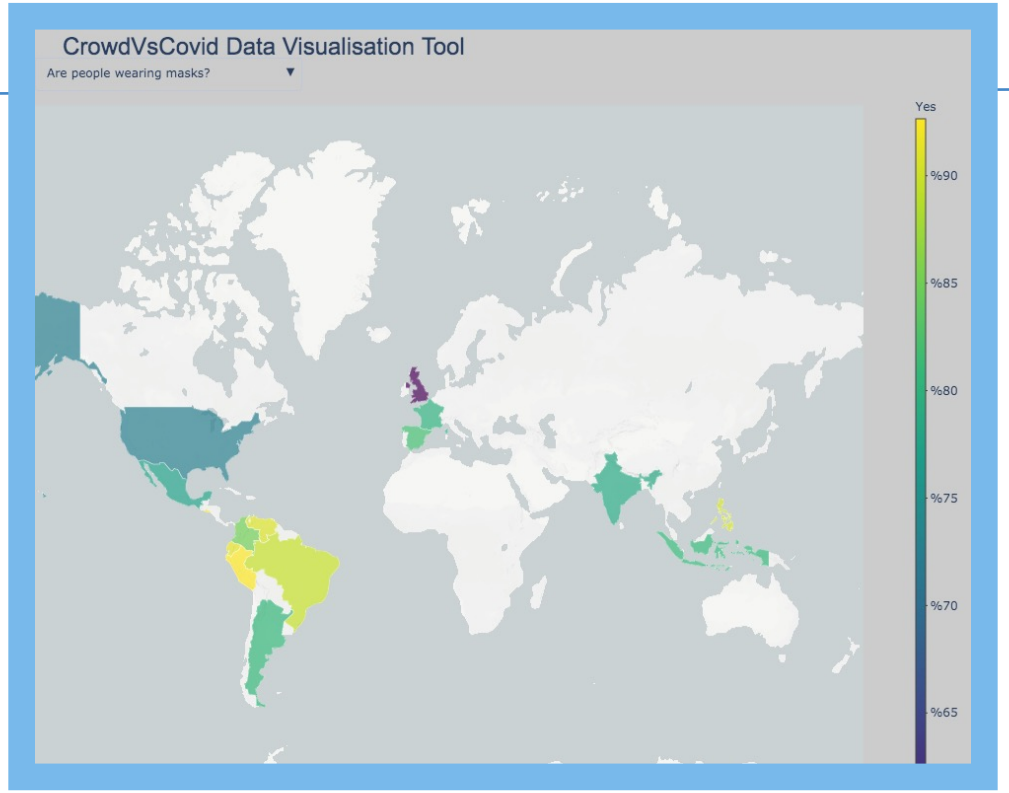
Crowd4SDG

- Based on native or augmented information


- Computed on filtered data at some stage(s)

Aggregation and descriptive statistics

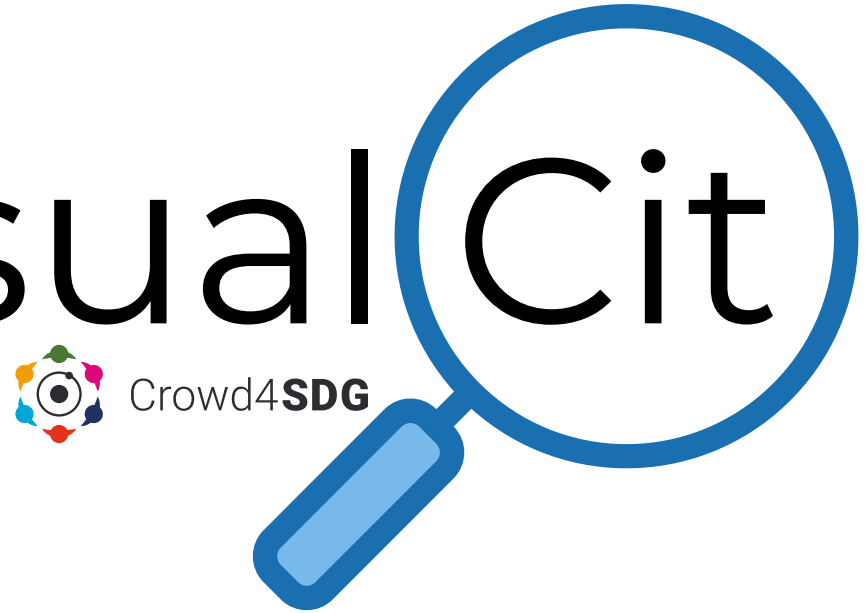
Crowd4SDG


# Visualisation on Maps

Crowd4SDG

# VisualCit




*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 872944*

https://pernici.faculty.polimi.it/crowd4sdgpolimi/