To what regards the data collection, we actually completed most of the tasks already at the last presentation; these were:
- The creation of a Mongo-Dataset to store the metadata of the downloaded pictures [CAMBIA]
- The start of the downloading procedure through the socials' API (FlickrApi and InstagramAPI) [CAMBIA]
- The saving of considered pictures into dropbox for the further analysis [CAMBIA]
- Finally, the insertion of the metadata into the db
It is to be noticed that the last two passages were performed together for each picture and only once sure that the picture hasn't been studied before. [CAMBIA]
The cleaning part was on the other hand still a WIP, especially the one regarding Instagram. [CAMBIA]

As the image shows, in the first version of the database, represented with a red line, a relevant increasing trend can be seen; this can be explained since, for 60 locations, all the available pictures were initially downloaded.
To prevent a huge number of results, we later decided to save the pictures according to some proportions (i.e. if the page had more than 20.000 pictures only 1 out of 20 was downloaded). On such a way we'd still be able to get meaningful results without overwhelming the dataset.
However, still a growth is visible in the light blue line. For justifying it, we identified a couple of reasons:[CAMBIA]

1) We Are Socials statistics says that in the last years Instagram has foreseen a considerable growth; this is even more noticeable considering that two years ago it wasn't even included in the top ones. So, our trend is reasonable according to this general growt. [CAMBIA]
2) More practically, it might just be a collection bias; Instaloader does not allow neither a research per year or coordinates. So, this means that the searching always starts from the most recent posts and only researches for Page, Hashtags and Location_ids are possible, where the latter are only very big numeric identifier. The only way we had to find those Id, with no list available, was iterating through the photos of well known pages regarding Turin and saving the new results when available.[CAMBIA]

Eventually, this lack of a global list, together with the fact that each user can create its own identifier and that in any case only discrete locations would have been obtained (very difficult to compare with the continuous distribution of Flickr) made us choosing to use this dataset only for the image processing part, without using it for the clustering phase.