V!RUS13

para tempos difíceis
Boas notícias
Good news
for hard times

ano 2016 year
semestre 2 semester

revista do nomads.usp I nomads.usp journal
issn:2175.974x I CC BY - NC
DOI 10.4237/ virus_journal

# THE IMAGE OF A DATA CITY

**Lev Manovich**
**Agustín Indaco**

**Como citar esse texto:** Manovich, L. and Indaco, A. The image of a data city. *V!RUS*, [e-journal] 13. Available at: <http://www.nomads.usp.br/virus/virus13/?sec=4&item=1&lang=en> [Accessed 00 Month 0000].

**Lev Manovich** is Ph.D. in Visual and Cultural Studies, Professor at City University of New York (CUNY), director of Software Studies Initiative. He studies cultural analytics, social computing, big data and society, digital humanities, digital art history, history and theory of media, software studies, data visualization.

**Agustín Indaco** is Master in Economy, researcher at Software Studies Initiative. He studies the intersection among applied microeconomics, health and big data, in addition to analysis of economic behavior through data obtained in social media.Social media content shared today in cities, such as Instagram images, their tags and descriptions, is the key form of contemporary urban life. It tells people where activities and locations that interest them are, and it allows them to share their urban experiences and self-representations. It creates an "image of a city" for both its residents and the outside world. One can argue that the identity of any city today is as much composed of the media content shared in that city on social networks as its infrastructure and economic activities. For these reasons, any analysis of city experience and self-representation needs to consider social media content shared in a given city.

**Keywords:** Big Data; Instagram; Gini index; New York; Social media.

Computational analysis of large numbers of user-generated photos and videos shared in particular areas can also help us to understand *how* people experience architecture and urban structures and *what* they do there. This can be done on any scale, from cities to the hyperlocal level of streets, buildings or parts of interiors. It is possible to compare the percentage of Instagram photos that show built environment in different cities, analyse which points of view are most popular for every landmark, and what emotions they evoke depending on time of the day. We can compare these patterns for residents and for tourists, for different genders, ages, and so on. In short, being able to analyse digital traces of what large number of people *do* in our built environments and how they *see* and *use* them can be very useful.

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
ano 2016 year | semestre 2 semester
revista do nomads.usp I nomads.usp journal
issn:2175.974x I CC BY - NC
DOI 10.4237/ virus_journal

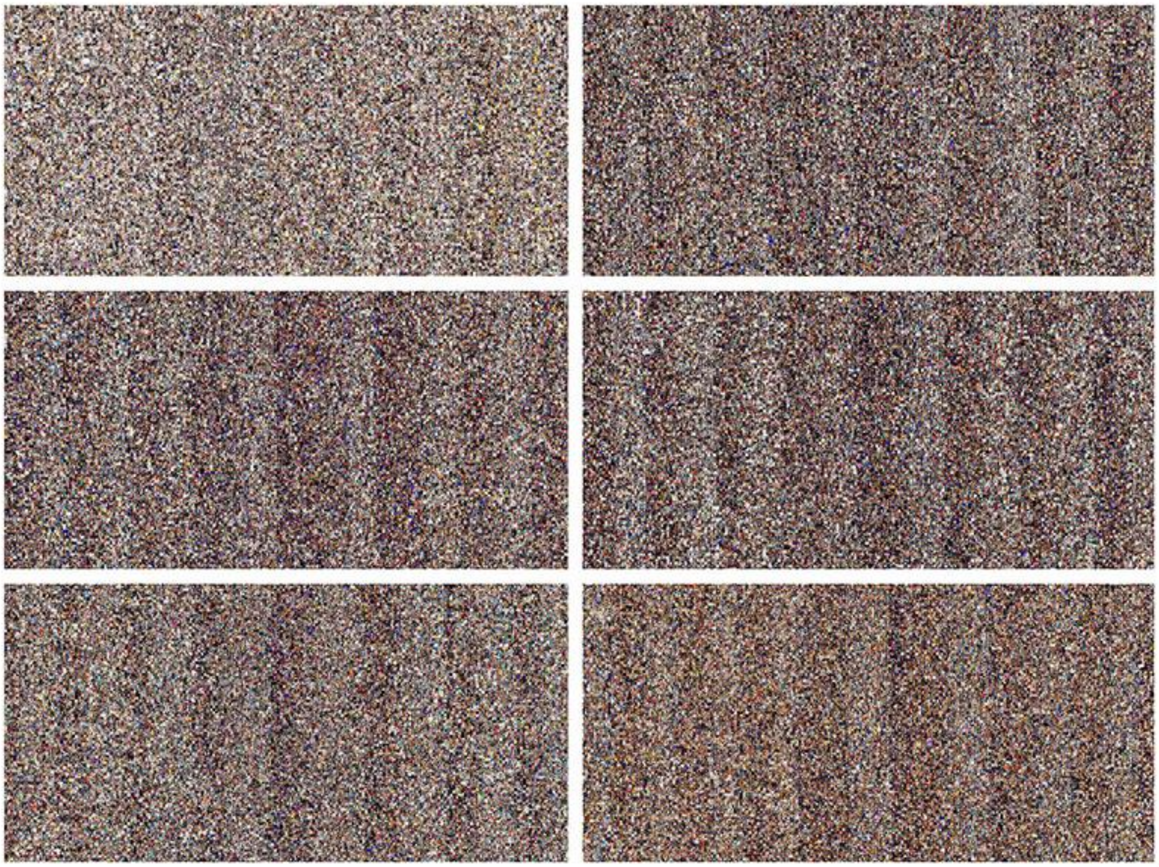

**Fig. 1:** Comparison of Instagram activity in different cities of Bangkok, Berlin, Moscow, New York, Sao Paolo and Tokyo (left to right and top to bottom). Each visualization shows 20,000 images shared consequently over one week in a given city. Source: Lev Manovich and Jay Chow, 2013-2016. Copyright: Software Studies Initiative.

In our Software Studies Lab (softwarestudies.com) located at University of California, San Diego and The Graduate Center, City University of New York we have been analysing over 16 million Instagram photos shared in 17 global cities starting in 2012. The research teams included data scientists, software developers, data visualization designers, media theorists, art historians, economists, and urban designers. Starting with a general comparison between 2.3 million images shared in 13 global cities (*Phototrails*, 2013, http://phototrails.net/*),* we consequently focused on more specific types of images, filtered by type of content: self-portraits (*Selfiecity*, 2014, http://selfiecity.net), a particular street (*On Broadway*, 2014, http://on-broadway.net), and a combination of a city area and a time period (the centre of Kiev during Maidan revolution of 2014 in http://www.the-everyday.net/). The illustrations for this essay present some of these projects.

While the lab's work shows how social media data can be useful for understanding the hyperlocal, it also reveals the limitations of this type of data. In many central urban areas social media has very high spatial and temporal resolution. For the *Inequaligram* project, the lab collected all 7,442,454 geo-coded Instagram photos publically shared in Manhattan during five months of March – August 2014. For example, in a single 30 m x 100 m area at Times Square, Instagram users shared 43,541 images. But in many other areas of Manhattan, people shared only a few dozen images during the same five

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
ano 2016 year | semestre 2 semester
revista do nomads.usp I nomads.usp journal
issn:2175.974x I CC BY - NC
DOI 10.4237/ virus_journal

months. Such low density in many parts of cities limits the usefulness of social media in understanding city life in such areas. Another limitation is demographic: for example, in many world cities only younger, well-educated people may post content. So while in some cases social media is a great resource to study hyperlocal locations, in other cases direct observation or surveys will be more useful. Therefore, large-scale computational urban social media analysis can only supplement - as opposed to replace other research methods in urban studies, design and architecture.

**Fig. 2:** *On Broadway* project combing an interactive installation and a website (http://on-broadway.nyc.) A screenshot from the interactive installation with a full zoom-out view showing the full length (13 miles) of Broadway street in Manhattan. The installation was shown at New York Public Library, December 2014 – January 2016. Source: Daniel Goddemeyer, Moritz Stefaner, Dominikus Baur, Lev Manovich, 2014. Copyright: Software Studies Initiative.

### *On Broadway*

The two most detailed studies of social media on a hyperlocal urban scale to date carried out by Software Studies Initiative are *On Broadway* and *Inequaligram* projects. Commissioned by New York Public Library, *On Broadway* (2015) was based on the original concept of media designer Daniel Goddemeyer. Moritz Stefaner was responsible for artistic direction and data visualization design, and Dominikus Baur for software development.

The project focused on a single very long street - part of Broadway that crosses all Manhattan (21 kilometers). The project team also wanted to include a slightly wider area than the street itself so we can capture the activities nearby. To define these areas, the researchers divided Broadway street into 30 meter-long segments, and then selected 100-meter wide rectangle areas around each segment centered on every point. The result was 713 identical 30 m x 100 m rectangles. The project visualizes and compares social media images and other data across these 713 areas.

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
ano 2016 year | semestre 2 semester
revista do nomads.usp I nomads.usp journal
issn:2175.974x I CC BY - NC
DOI 10.4237/ virus_journal

The main goal of the project was to construct a novel mechanism for navigating a "data city" consisting from many layers of images and data. We asked ourselves if there was a different way to visualize urban structures and activities besides maps, graphs, and numbers. The result of many explorations is a visually rich, image-centric interface, where numbers play only a secondary role, and no maps are used. This interface proposes a new visual metaphor for thinking about the city: a vertical stack of image and data layers. There are 13 such layers in the project, all aligned to locations along Broadway. They include images shared along Broadway on Instagram and Twitter, images from Google Street View, Foursquare check-ins, taxi rides, and selected economic and social indicators from the U.S. Census. Overall, we used over 30 million data points and images to represent activities along a single street.

As you move along the representation of a street, you see a selection of Instagram photos from each area, left, right, and top Google Street View images and extracted top colours from these image sources. You also see the average number of taxi pickups and drop-offs, Twitter posts with images, and average family income for the parts of the city crossed by Broadway. To help with navigation, we added additional layers showing names of Manhattan neighbourhoods crossed by Broadway, cross-streets and landmarks.

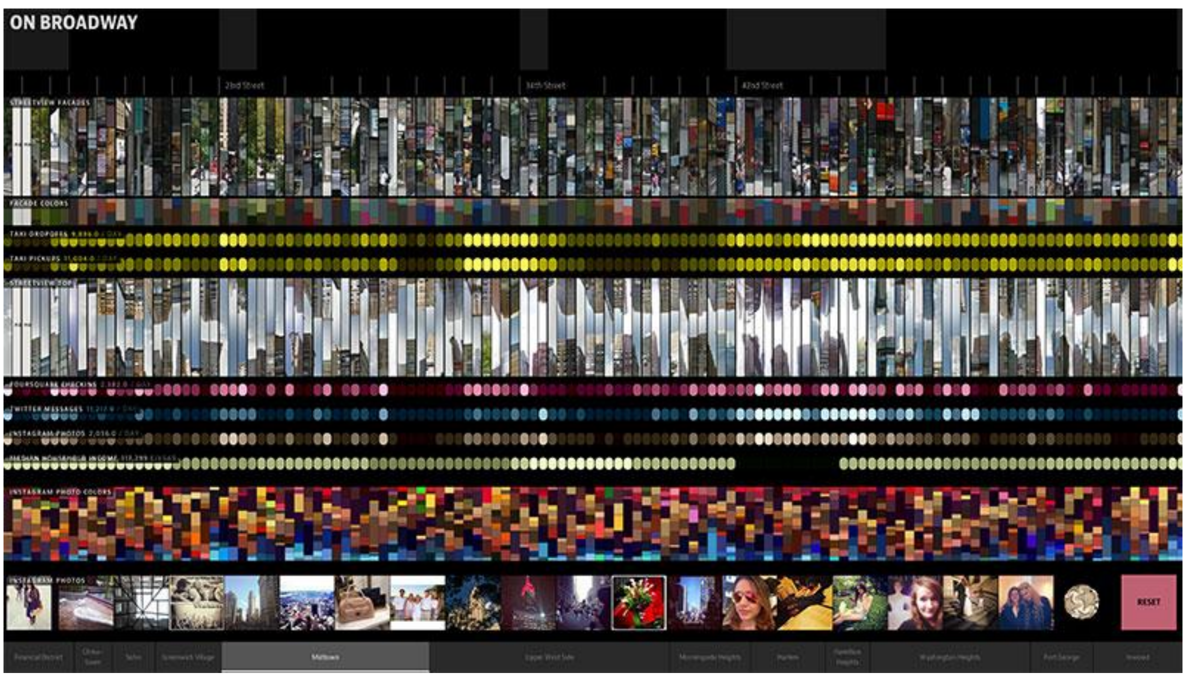


**Fig. 3:** Daniel Goddemeyer, Moritz Stefaner, Dominikus Baur, Lev Manovich, 2014. *On Broadway*. A screenshot from the interactive installation. Neighbourhood-level zoom view showing midtown area in Manhattan. The installation was shown at New York Public Library, December 2014 – January 2016. Copyright: Software Studies Initiative.

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
ano 2016 year | semestre 2 semester
revista do nomads.usp | nomads.usp journal
issn:2175.974x | CC BY - NC
DOI 10.4237/ virus_journal

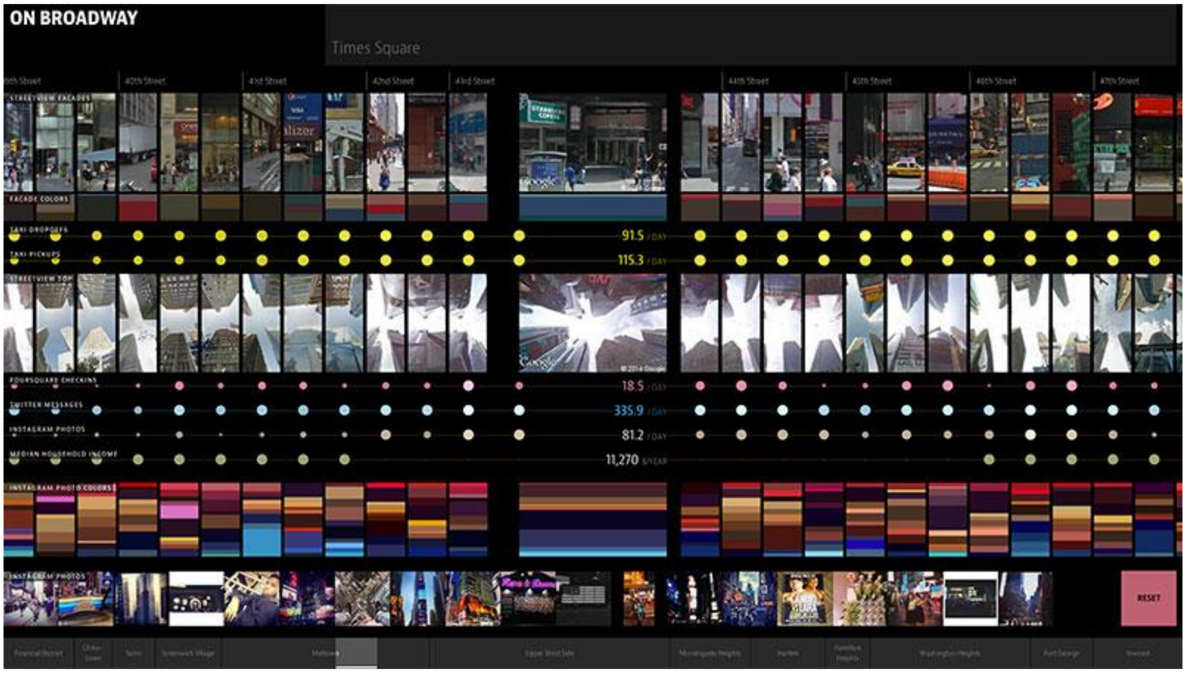


**Fig. 4:** Daniel Goddemeyer. Moritz Stefaner, Dominikus Baur, Lev Manovich, 2014. *On Broadway*. A screenshot from the interactive installation. Block-level zoom view cantered on Time Square area in Manhattan. Copyright: Software Studies Initiative.


**Fig. 5:** Daniel Goddemeyer, Moritz Stefaner, Dominikus Baur, Lev Manovich, 2014. *On Broadway*. A graph comparing the data layers used to represent Broadway street in the project. Broadway street is projected

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
ano 2016 year | semestre 2 semester
revista do nomads.usp I nomads.usp journal
issn:2175.974x I CC BY - NC
DOI 10.4237/ virus_journal

onto horizontal axis (south to north becomes left to right). The height of a graph at every location corresponds to volume of a particular data layer at this location. Copyright: Software Studies Initiative.

This multi-layered Broadway "corridor" can be explored on many scales. In zoomed out view, you see all 21 kilometers of the street. To do this, we are displaying narrow vertical slices of every Google Street photo. When you start zooming in, the slices become wider. Finally, in a complete zoomed in view, the image of the currently selected area is shown in full size. All visuals in all layers and numbers showing aggregated activity are instantly updated when a user moves right or left, or changes zoom level.

When the project team was exploring all data layers along Broadway, it found that volumes of all data layers are strongly correlated. Informally this can be seen on the graph that plots volumes of all variables we looked at: the variables go up and down together. How is it possible to interpret this "correlated city?" The data suggest that social inequality and digital divide are now joined by a social media divide that is even more extreme. In affluent areas, people make more money, take taxis, and post more images on Instagram and Twitter. In poor areas, people make less money, rarely use taxis, and post much fewer images on social networks.

## *Inequaligram*

The members of the lab decided to further explore these connections. The result is a new project called *Inequaligram* (2016) created by the authors of this essay. It analyzes characteristics of Instagram posts and socio-economic indicators reported by the U.S. Census across Manhattan.

U.S. Census reports aggregate socio-economic characteristics of populations using a type of division called "tracts." There are 287 census tracts in Manhattan. The average population of each tract is 3,000-4,000 people and its average size is 0.36 square km. The project uses these tract areas to compare patterns in Instagram sharing and indicators such as income and unemployment rate.

The project team chose Instagram for this analysis because it has the strongest geographic and spatial identity among all top social media services. While tweets and Facebook posts can also have geo-coordinates and talk about the local events around the user at the moment of posting, Instagram images often directly capture these events and show users in particular places. And since Instagram posts contain an image or a video, date and time metadata, descriptions, and hashtags, this allows us to study collective representations of city life along these separate dimensions. For example, we can compare the number of images shared between areas, presence of different subjects in these images, most popular and most unique hashtags, how people are dressed and so on. These and many other characteristics can be extracted automatically from Instagram posts using data science techniques available in open source software.

Social media content shared in a given area may combine contributions from different kinds of users: people who reside in this area, people who live in different parts of the city or in suburbs but spend significant time in this area for work during weekdays; international or domestic tourists visiting a city; companies located in this area, and so on. Together, the content shared by all these users create a collective "voice" of a

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
a n o 2016 y e a r | semestre 2 semester
revista do nomads.usp I nomads.usp journal
issn:2175.974x I CC BY - NC
DOI 10.4237/ virus_journal

particular area of a city. A city as a whole can be compared to a chorus of all these voices although, of course, they are not necessary performing the same composition. Applying the concept of inequality to a collection of these urban voices can give us new ways of understanding a city, and provide an additional metric for comparing numerous cities around the world.

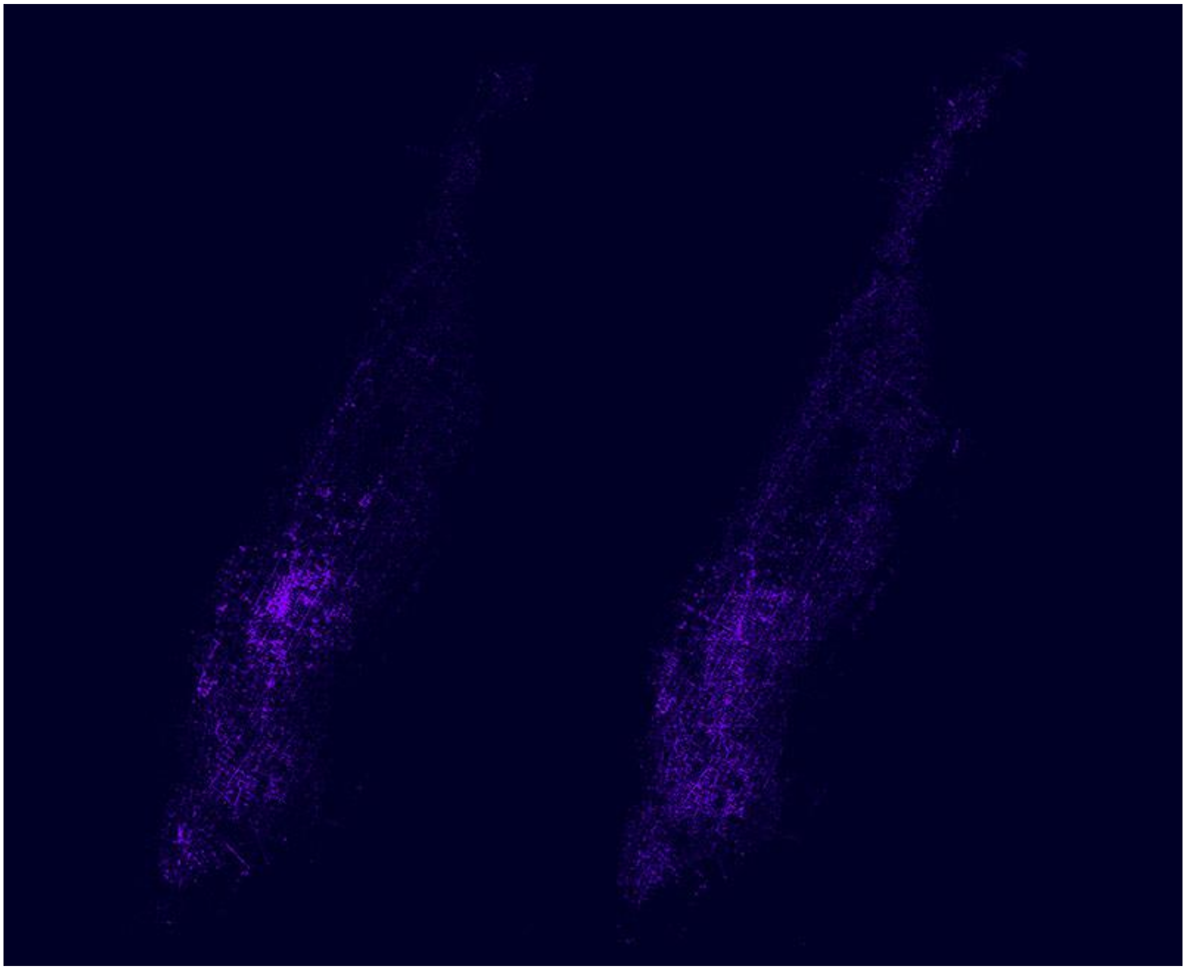

**Fig. 6**: *Inequaligram*. Locations of Instagram images shared by NYC visitors (left) and locals (right). Each map uses a 100,000 random image sample. They are drawn from the larger set of 7,442,454 geo-tagged images publicly shared in Manhattan during 3/2014-7/2014. Source: Agustin Indaco and Lev Manovich, 2016. Copyright: Software Studies Initiative.

In contrast to other social media services, image and location driven by Instagram create an "image of a city" for both locals and visitors. Therefore, we need to understand what such collective representations contain and how their characteristics are related to both a city's architectural structure (for example, presence of tourist landmarks) and socio-economic social structure (for example, the locations of rich/poor areas).

Urban planners and architects know how to map cities' physical structures, but what are the most informative ways for them to map and analyze social media? In a city like New York, people share a very large number of Instagram images in some areas and very few in others. The images shared in some areas may also contain more hashtags

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
ano 2016 year | semestre 2 semester
revista do nomads.usp | nomads.usp journal
issn:2175.974x | CC BY - NC
DOI 10.4237/ virus_journal

and descriptions that talk about local architecture than in other areas. When we plot such characteristics of users' posts using their geo-locations, we see that their spatial distributions are very uneven.

To be able to quantify exactly how uneven these distributions are, the *Inequaligram* team developed a new concept of "social media inequality." This concept allows us to quantitatively compare spatial patterns in relevant social media activity between parts of a city, a number of cities, or any other spatial areas. The team defined this concept using an analogy with the concept of economic inequality. Economic inequality indicates how some economic characteristic or material resource, such as income, wealth or consumption is distributed in a city, country or between countries. Accordingly, social media inequality indicates how some characteristic of shared social media content is distributed between geographic areas. Examples of such characteristics are the number of photos shared by all users of a social network such as Instagram in a given city area, numbers of hashtags, and numbers of unique hashtags.

To compensate for the differences in the geographic size of tracts, Instagram data was normalized by tract size. The project also used the dates of shared images to estimate if a particular user lives in Manhattan or is only visiting. Data from the U.S. Office of Travel and Tourism Industries indicates that the average visitor stays 10.5 days in New York City. We decided to use a slightly larger 12-day period, and consider a user a "visitor" if she posted all her photos within a single 12-day period out of the total five months of our data collection. On the other hand, if a user shared a minimum of two photos within any interval larger than 12 days, we consider this person a "local." Although this very simple method is not precise, analysis of the data show that it does effectively differentiates captures between these two groups. Our dataset contains 5,918,408 million images from 366,539 unique Instagram accounts of local residents, and 1,524,046 images from 505,345 accounts that belong to visitors.


**Fig. 7**: *Inequaligram*. Hourly proportions of images shared by locals in selected Manhattan neighbourhoods below the 59th street. The graph uses time stamps of 5,918,408 million images shared by 366,539 local residents. Source: Agustin Indaco and Lev Manovich, 2016. Copyright: Software Studies Initiative.

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
ano 2016 year | semestre 2 semester
revista do nomads.usp | nomads.usp journal
issn:2175.974x | CC BY - NC
DOI 10.4237/ virus_journal

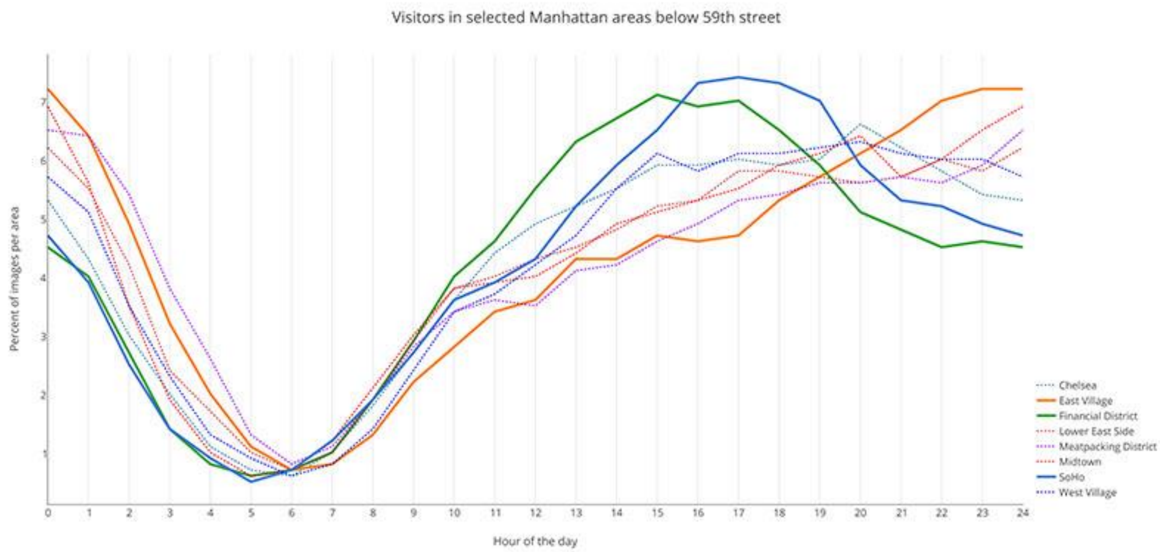


**Fig. 8:** *Inequaligram*. Hourly proportions of images shared by visitors in selected Manhattan neighbourhoods below the 59th street. The graph uses time stamps of 1,524,046 images shared by 505,345 visitors. Source: Agustin Indaco and Lev Manovich, 2016. Copyright: Software Studies Initiative.


**Fig. 9:** *Inequaligram*. Hourly proportions of images shared by locals in selected Manhattan neighborhoods above the 59th street. The graph uses time stamps of 5,918,408 million images shared by 366,539 local residents. Source: Agustin Indaco and Lev Manovich, 2016. Copyright: Software Studies Initiative.

To compare social media inequality across Manhattan for these two groups, *Inequaligram* decided to use the most popular measure of economic inequality – the Gini index. This is the same measurement used in most discussions of income and wealth inequality in both economics and in popular press. In the case of Instagram, if people were to share exactly the same number of images each in each city tract, this means complete equality, and Gini index = 0. If, on the other hand, people were to

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
ano 2016 year | semestre 2 semester
revista do nomads.usp I nomads.usp journal
issn:2175.974x I CC BY - NC
DOI 10.4237/ virus_journal

share all images in only one tract, and nothing in all other tracts, this means complete inequality, and Gini index = 1.

*Inequaligram* found that Gini index for the number of images shared in Manhattan between all tracts is 0.494 for locals, and 0.669 for visitors. For the total numbers of hashtags, the index is even higher: 0.514 for locals, and 0.678 for visitors. To put this in context, Instagram inequality for numbers of visitors' images in Manhattan (Gini = 0.669) is larger than income inequality in the most unequal country in the world (Seychelles where Gini = 0.658). Social media shared by locals has a Gini coefficient similar to countries that rank between 25 and 30 in the list of countries by income inequality. These are countries like Costa Rica (0.486), Mexico (0.481) and Ecuador (0.466).

What drives high inequality of Instagram sharing between parts of Manhattan? In the case of visitors, they share most images in midtown Manhattan (big shopping and hotels area), around famous landmarks such as Times Square and the Flatiron Building, and in the evening dining and drinking areas like East Village and Lower East Side. In the case of locals, our analysis suggests that differences in their social media activity among parts of a city are to a large extent driven by commuting patterns. During work hours on weekdays the residents of less prosperous areas such as parts of Manhattan above $100^{th}$ street work in more prosperous parts of the city - areas below 100th street, and particularly in Midtown. This is where they share images on Instagram during the day, so their shares get added to these areas.

Looking at inequality patterns in Instagram shares of locals and visitors together, *Inequaligram* found that the areas of Manhattan below 100th street with most businesses are also the ones that are the most popular among visitors. Thus, we have the effect of *double amplification* – social media contributions by affluent residents from these areas get amplified by the contributions of people who commute there for work from other parts of Manhattan, and also by contributions from out-of-city visitors. Comparing social media statistics with Census indicators for tracts in Manhattan, we find that the inequality of numbers of Instagram images between tracts is *bigger* than inequalities in levels of income, rent, and unemployment. Gini indexes are 0.32 (median income), 0.22 (median rent), 0.35 (unemployment rate), and 0.49 (numbers of Instagram images shared by local residents). This is a very interesting and original result.

V!RUS13
para tempos difíceis
Boas notícias
Good news
for hard times
a n o 2016 y e a r | semestre 2 semester
revista do nomads.usp I nomads.usp journal
issn:2175.974x I CC BY - NC
DOI 10.4237/ virus_journal

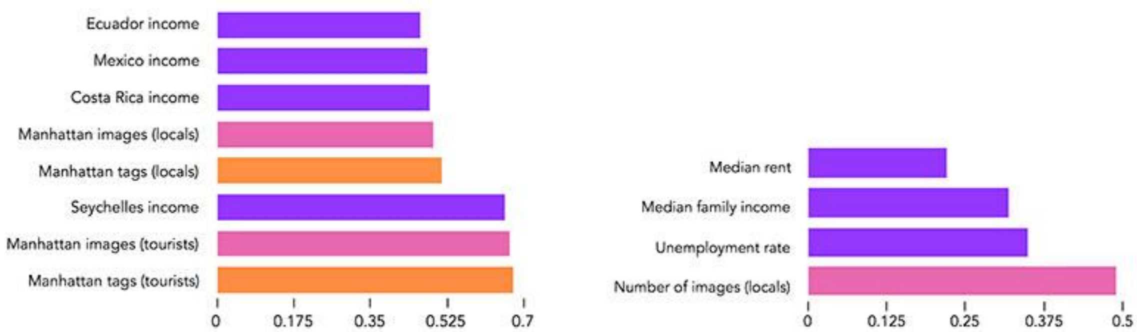


**Fig. 10**: *Inequaligram*. Left: Gini inequality measure for numbers of Instagram images and tags shared in Manhattan compared to income inequality measures in selected countries. Right: Gini inequality measures for Instagram images shared by locals in 287 Manhattan tracts and selected Census economic indicators (rent, income, unemployment) for the same tracts. "Tracts" are spatial divisions used by U.S. Census in reporting surveys results. Gini measures for economic indicators are calculated using 2014 Census data. Source: Agustin Indaco and Lev Manovich, 2016. Copyright: Software Studies Initiative.

## Studying the Urban Life in the Data Era

There are many analytical possibilities that social media's big data offers to urban researchers and practicing urbanists and architects that can be explored besides those discussed here. By downloading, analyzing, and visualizing user-shared photos, along with their tags, descriptions, time stamps and geo-coordinates, the Software Studies Lab researchers have pieced together a collective "image of a city" and been able to see how it changes over time. The concept of social media inequality allows us to measure how this image changes from area to area, and also compare such images at arbitrary spatial scales.

A thorough analysis of cities and city life in the 21$^{st}$ century certainly should contain more layers than social media alone. And yet, as we suggest in this essay, the social media layer plays a very important role because it filters the city in particular ways, highlighting some locations and making others invisible. Social media data allows us to create new representations and new concepts that help us understand cities and city life in new ways. As *On Broadway* shows, we can construct new visual representations of cities that portray urban behavior and media using many scales and layers of data. And as *Inequaligram* shows, social media data also allows us to produce new metrics for understanding city life and comparing cities across the world. We believe that such new approaches will supplement other existing research methods in urban studies, architecture, media studies, and social sciences and will shape the way we understand urban life in the decades to come.