Big Data:

Qué es, qué no es, y cómo aplicarlo

Agustín Indaco
aindaco@gmail.com

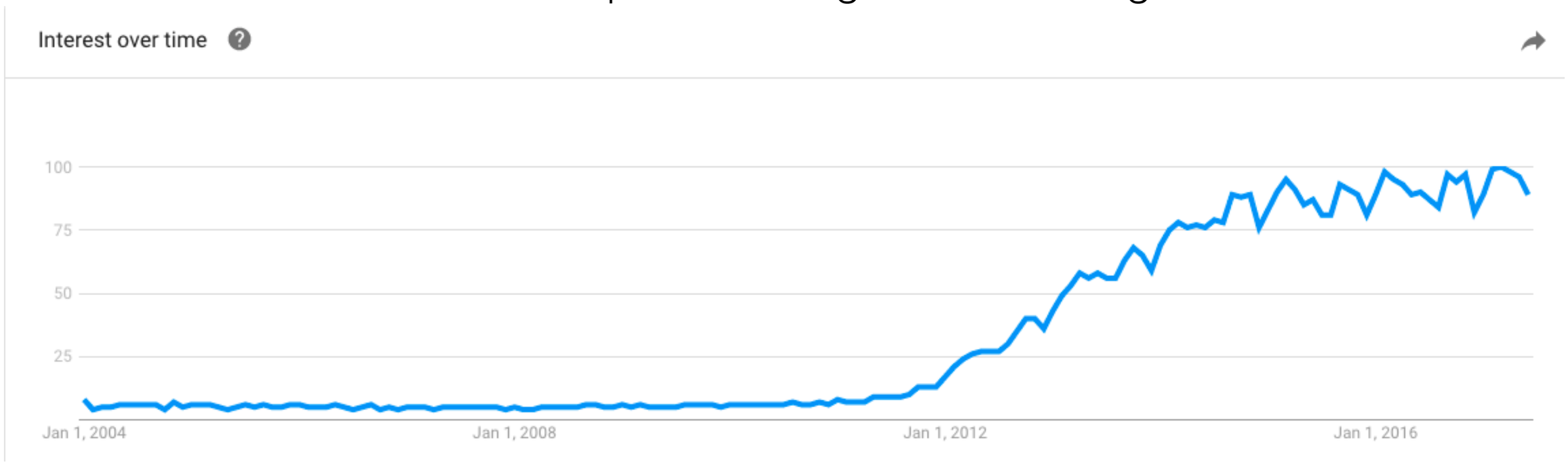
29 de junio, 2017

Jóvenes CAI

Big Data: la moda

La popularidad de Big Data como término parece haberse disparado en 2012 y se ha mantenido elevado.

Evolución de búsqueda de ‘Big Data’ en Google



Fuente: https://trends.google.com/trends/explore?date=all&q=big%20data

¿Qué es Big Data?

- Simplemente quiere decir que hay muchos datos.
- La magnitud puede estar en la ‘transversalidad’ o en la longitud de los datos.
- ¿Cuánto es muchos? Depende de qué definición usamos. En general, se considera que hablamos de datos que no podés guardar en la memoria de tu PC (datos > RAM).

¿Qué es Big Data?

- Más allá del tamaño en sí de los datos, la revolución está en el hecho de que nos permite estudiar cuestiones que antes eran imposibles e incluso resolver problemas que antes no podíamos.

- La revolución está en la originalidad de los datos: compras, movilidad, películas, ratings, etc.

Big data: el debate

Fin de la teoría: “¿Quién sabe por qué la gente hace lo que hace? El punto es que lo hacen, y que podemos rastrearlo y medirlo con una precisión antes impensable. Con suficientes datos los números hablan por sí mismos.” *Chris Anderson, Wired, 23/6/2008*

Problemas de ‘Small data’ en Big data: “Hay muchos problemas de small data que ocurren usando big data. Éstos no desaparecen al tener muchos datos, sino que empeoran.” *David Spiegelhalter, Financial Times, 28/3/2014*

Lo que Big Data no soluciona
(o incluso empeora)

- **No tenemos todos los datos:** se puede creer (equivocadamente) que con Big Data tenemos acceso a todos los datos, pero no es así.

- **Sesgo:** datos que surgen de Big Data comúnmente son no estructurados (a diferencia de una muestra). La falta de estructura incrementa el riesgo de sesgo.

- **La ley de grandes números no ayuda necesariamente:** si la muestra es sesgada, más observaciones no nos llevan a la media poblacional.

Predicción: un problema de sesgo vs varianza

Predicción: en la mayoría de los casos se usa Big Data para predecir.

$$Y = f(X) + u$$

Queremos predecir Y en base a datos sobre X, sin observar u y sin saber la función f(.)

Predicción: un problema de sesgo vs varianza

Predicción: Y = f(X) + u

Lo que queremos es reducir el Error Cuadrático Medio (ECM):

$$E\left[(Y - \hat{f}(X)^2)\right]$$

ECM = sesgo^2 + varianza

Sesgo: la diferencia entre el valor esperado y el valor del parámetro poblacional que estima.

Varianza: medida de dispersión de dicha variable respecto a su media.

Predicción: un problema de sesgo vs varianza

Por lo general los **modelos más complejos suelen tener sesgos más bajos**, pero la **varianza de su predicción es más elevada** por replicar el ruido de la muestra de 'entrenamiento'. Esto se llama 'overfitting'.

Modelos más simples suelen tener más sesgo pero pueden reducir la varianza de la predicción. Esto se suele llamar 'underfitting'.

Se suele elegir el modelo con menor ECM.

Muestra vs Big Data

Caso Gallup (1936):
Predecir elecciones EEUU: la revista The Literary Digest envió sobres a 10millones de habitantes preguntando a quién votarían (1/4 del electorado). Recibieron 2.5millones de respuestas con el 55% favoreciendo a Landon y 41% a Roosevelt (el resto a otro candidato).

Al mismo tiempo Martin Gallup hizo una encuesta a 3mil habitantes, preocupándose porque la muestra fuera representativa (y sabiendo que con 3mil, el margen de error iba a ser lo suficientemente pequeño para poder predecir). Su predicción fue muy cercana al resultado final: ganó Roosevelt (61% vs 37%).

Muestra vs Big Data

Caso Gallup (1936): ¿Por qué?

- Más datos de lo mismo no necesariamente agregan información
- En el caso de la revista, no sabían estructura de respuestas: quién había respondido (jóvenes, clase alta, hombres…?)
- El error muestral se reduce con ‘pocas’ observaciones (n=2,500 para estimar con 2% margen error en elecciones nacionales de Argentina).
- Para pensar: ¿las evaluaciones de un restaurant en Yelp es representativa? ¿Y la opinión de los que averiguaron acerca del restaurant pero decidieron ir a otro?

Muestra vs Big Data

Pero ‘Big Data’ nos permite:

- Saber **más características acerca de cada observación.** Encuestas suelen preguntar edad y género, pero muchas veces Big Data nos da más información: gustos, lugar de trabajo, red de contactos, etc. **Transversalidad de datos.**

- Al tener tantos datos, podemos particionar la población de manera más detallada y **sacar predicciones de grupos bien acotados** (ej.: mujeres ingenieras de entre 25-35 años que se radican en Buenos Aires). **Longitud de datos.**

Revolución del Big Data

La gran novedad del Big Data es que **nos permite obtener información que antes no teníamos**: transacciones, movimiento de personas, gustos, preferencias, etc. Y al tener muchas observaciones, podemos sacar conclusiones estadísticamente significativas acerca de grupos con características bien detalladas.

Esto **nos permite responder preguntas que antes eran imposibles de responder** y además encontrar un nuevo enfoque a preguntas viejas.

Big Data: Redes sociales

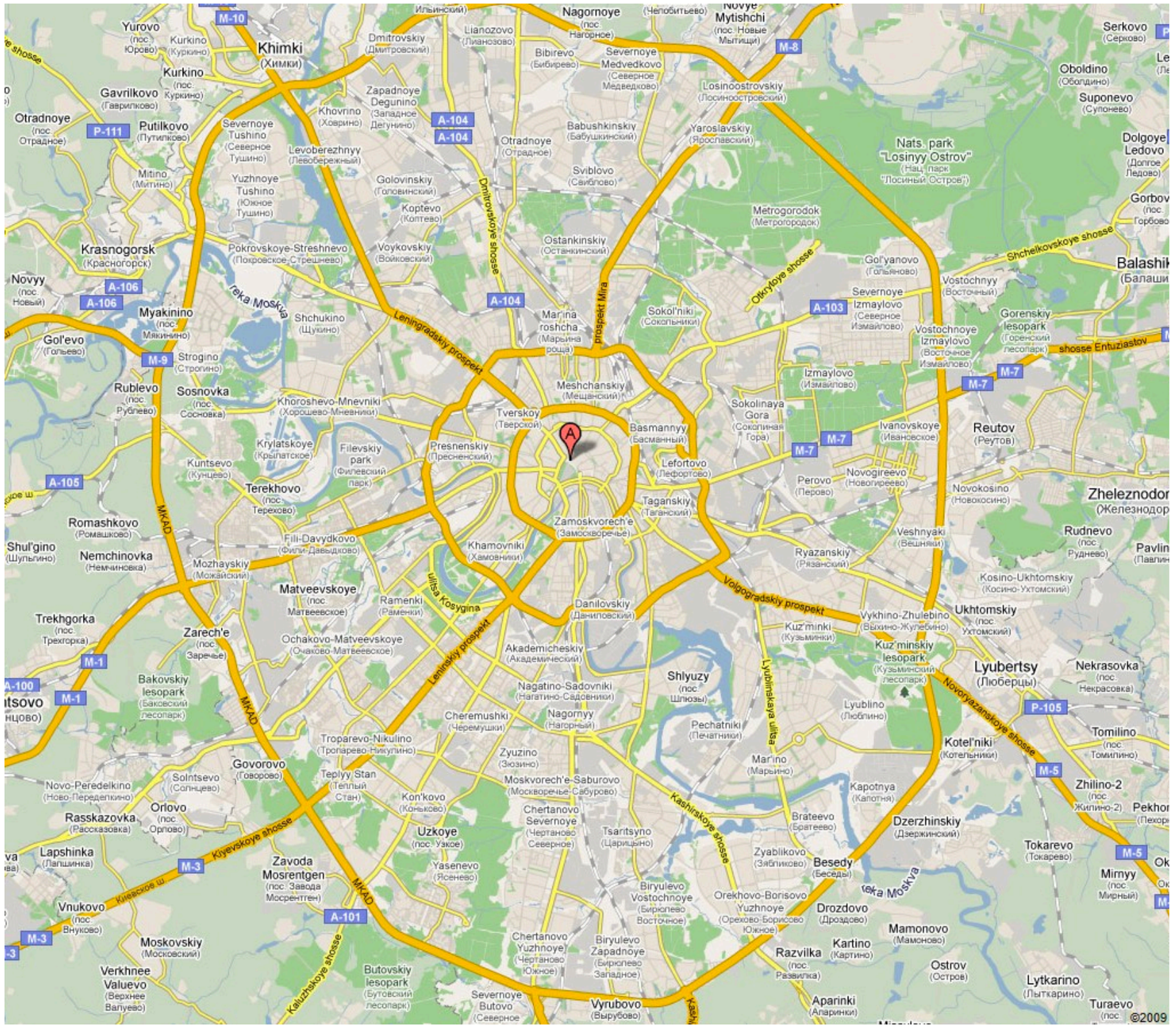
Una gran fuente de los datos novedosos a los que podemos acceder hoy provienen de las redes sociales.

Problemas:
Sesgo 1: los usuarios de las distintas redes sociales claramente **no son una muestra representativa de la población.**

Sesgo 2: incluso dentro de los usuarios de una red social, **algunos usuarios generan mucho más contenido que otros.** Por lo que las observaciones dentro de una plataforma pueden estar sesgados hacia las actividades de un subgrupo en particular.

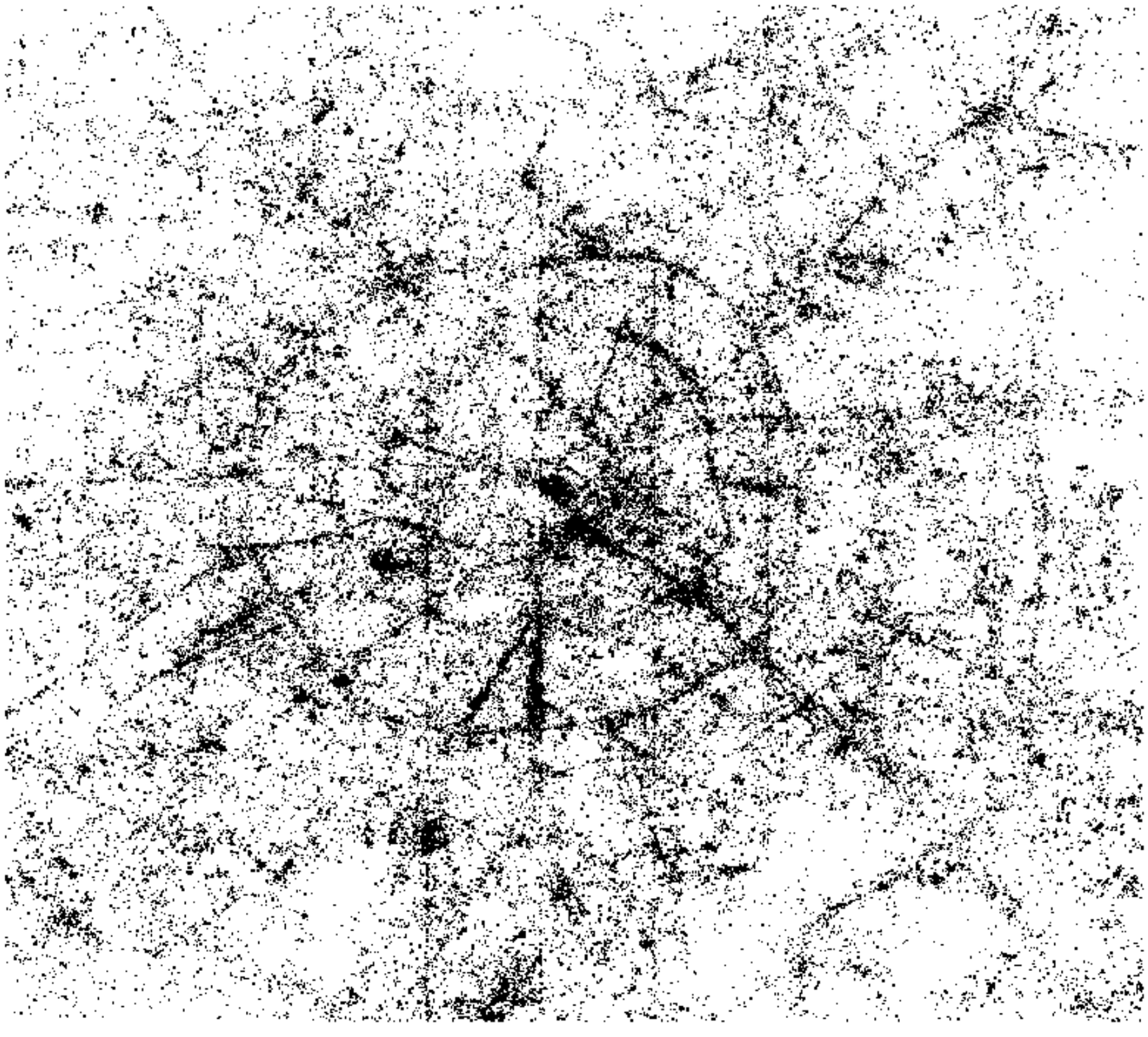
¿Hay patrones en los datos?

Mapa de Moscú



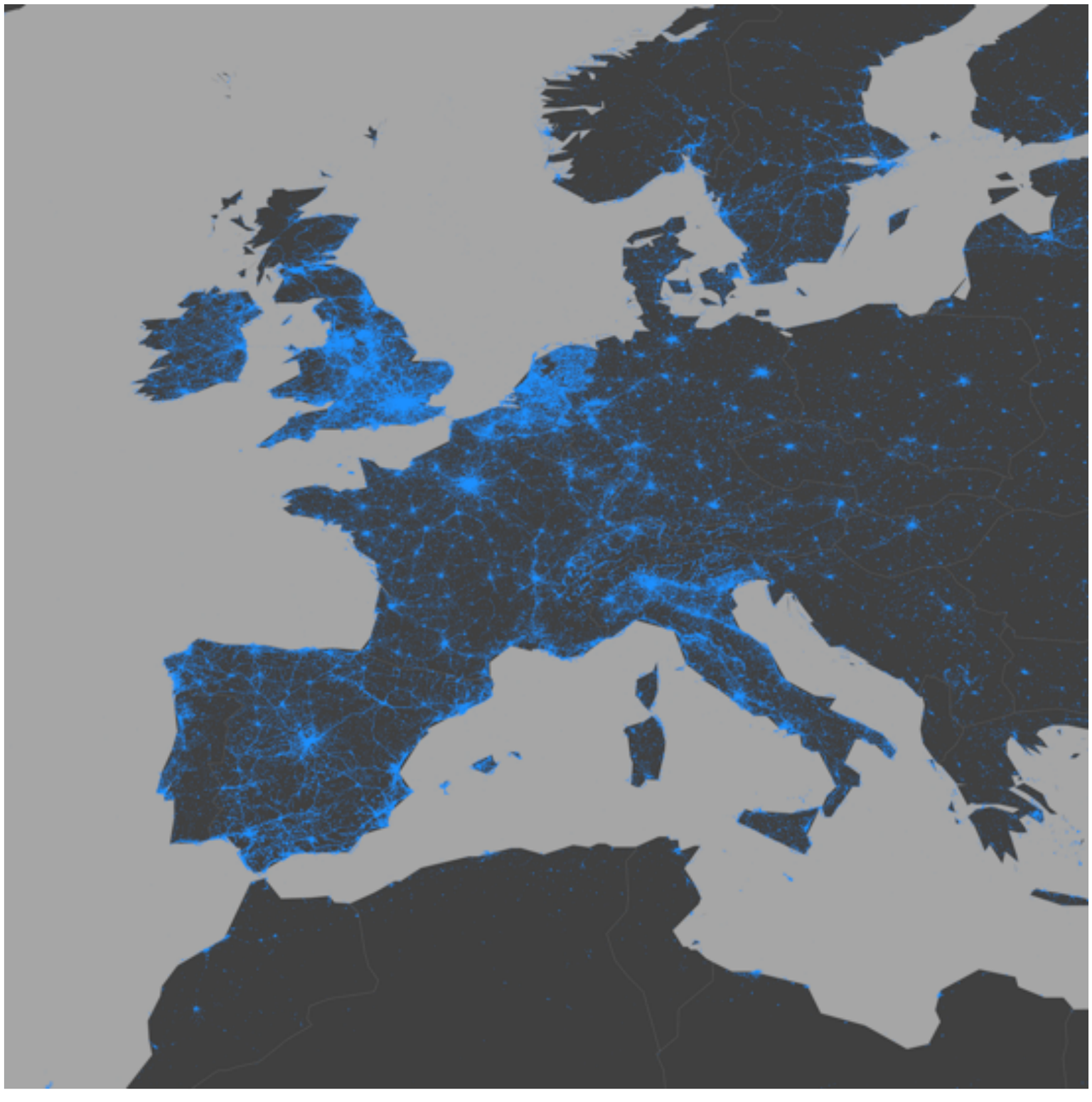
¿Hay patrones en los datos?

Cada punto representa un Tweet enviado en Moscú en 2012
(n=100mil)



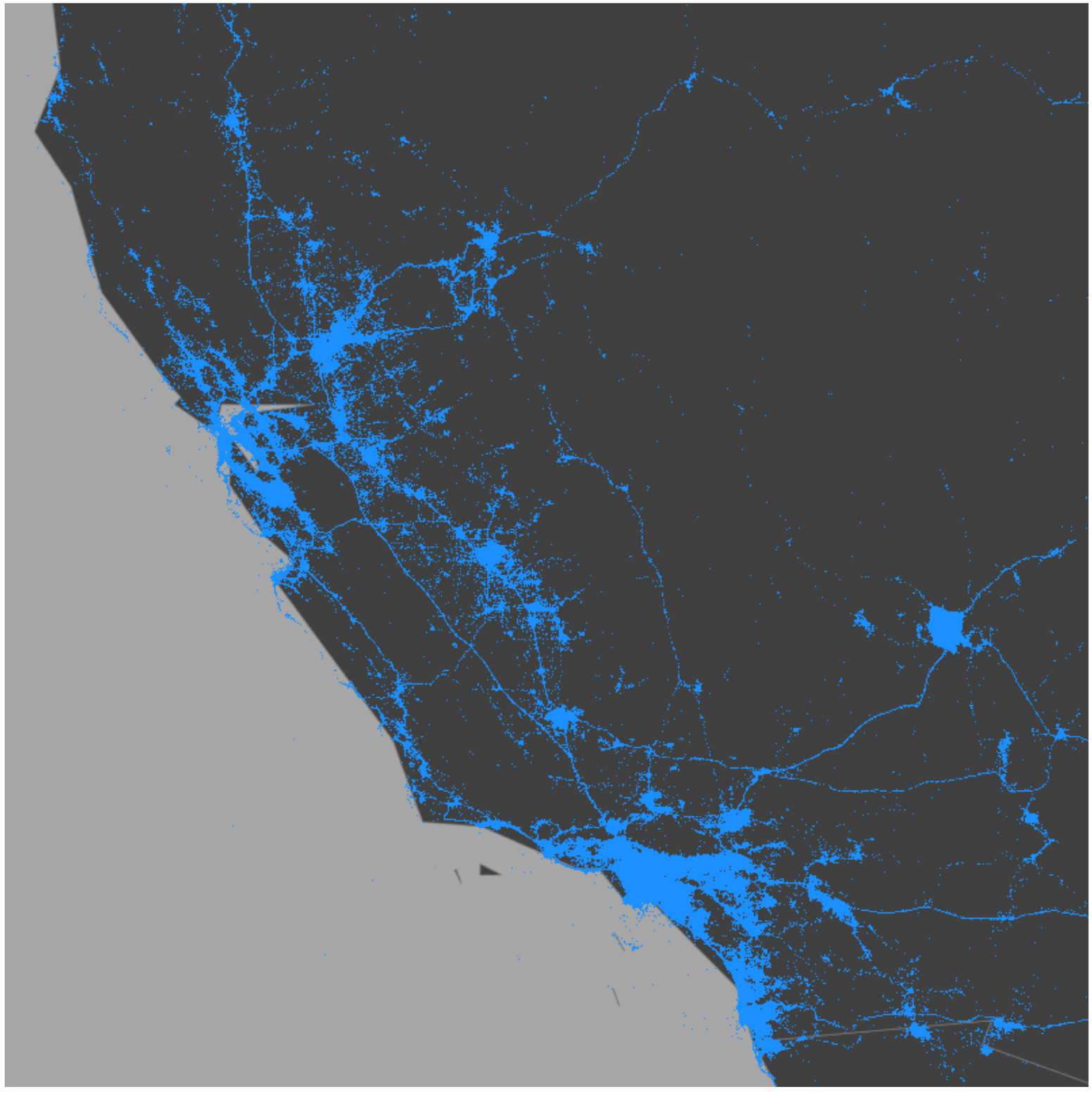
¿Hay patrones en los datos?

Tweets en Europa



¿Hay patrones en los datos?

Costa oeste de EE.UU.



Inequaligram: estudiando estructuras urbanas con Instagram

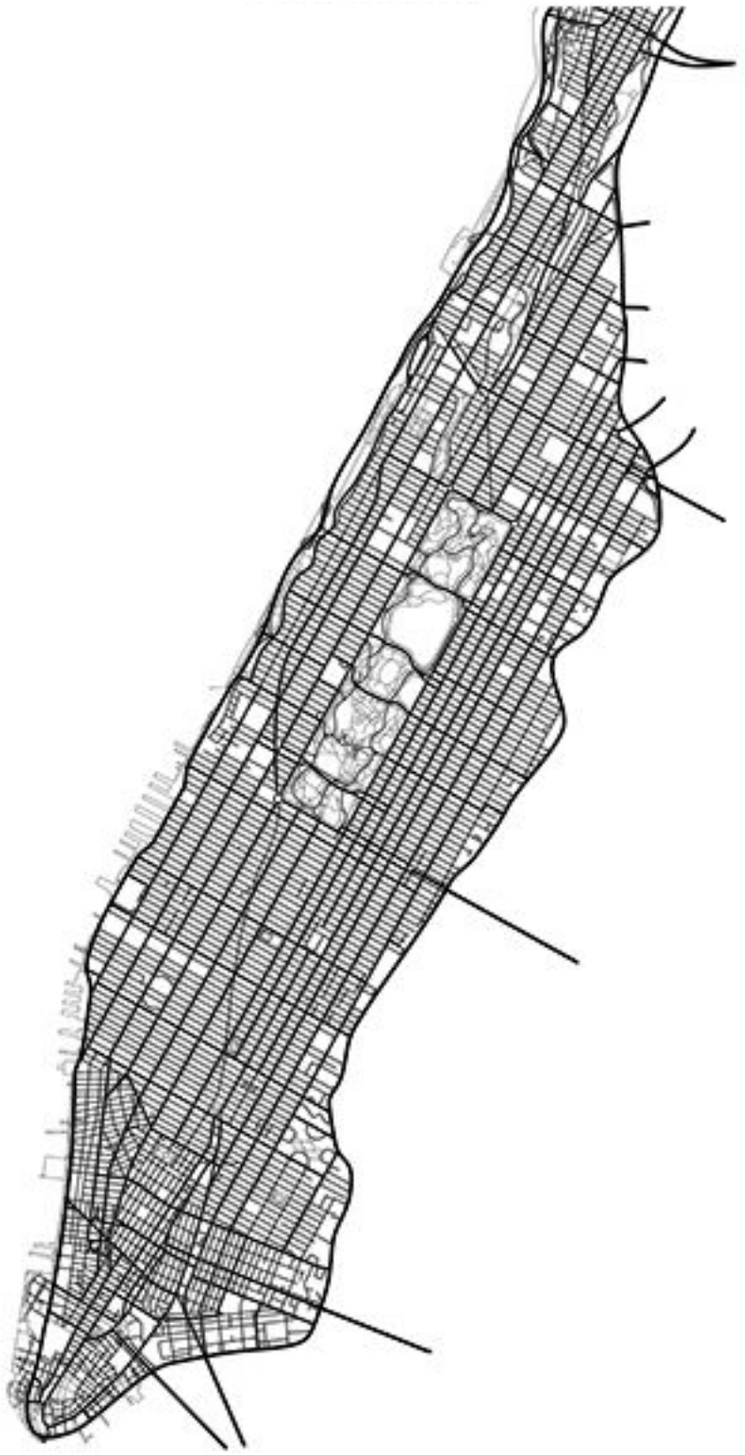
Inequaligram

- Analizamos **7.5 millones de fotos** subidas a Instagram en NY entre marzo-julio 2014

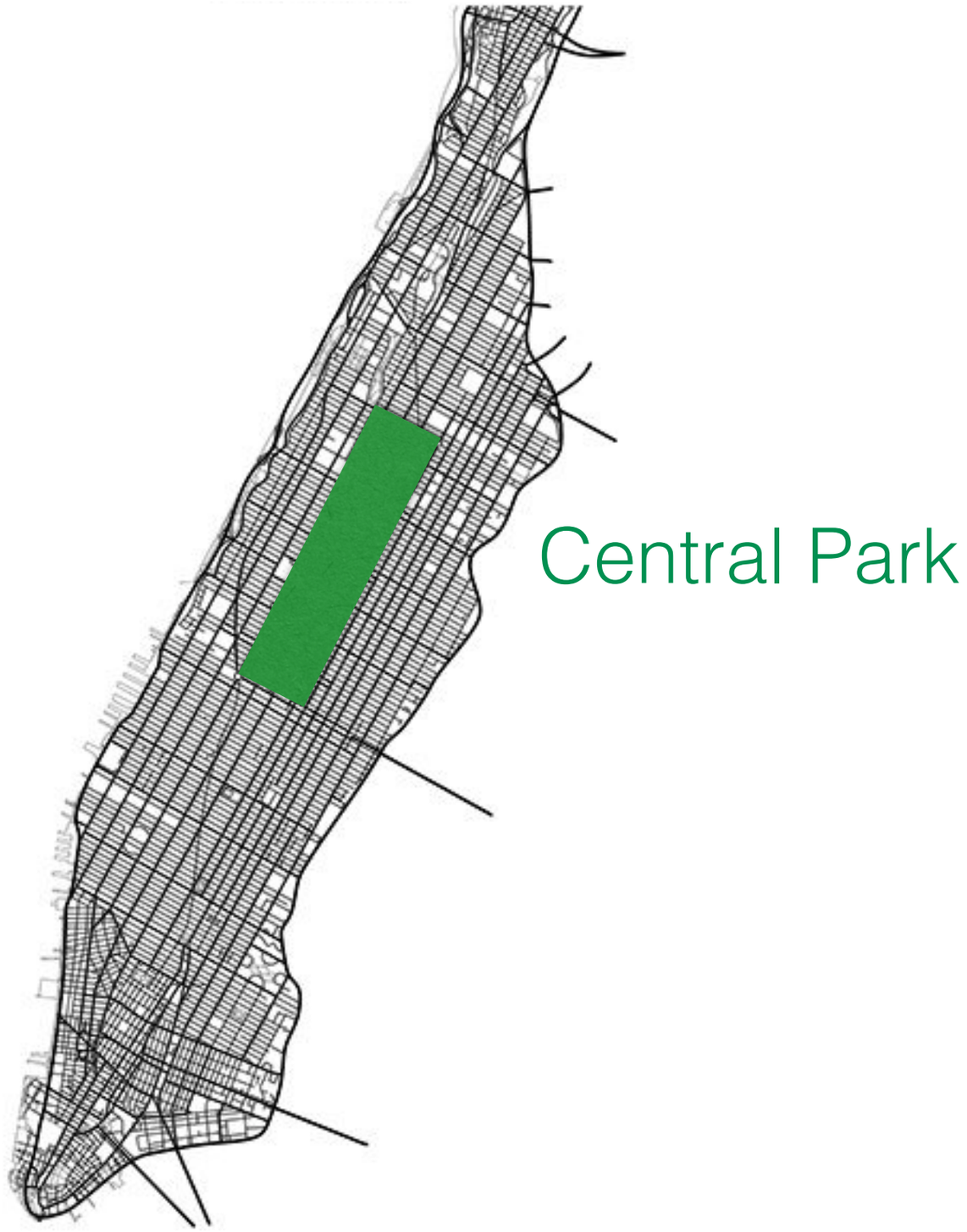
- Agrupamos imágenes en **zonas censales** (287 total en Manhattan)

- Dividimos muestra entre usuarios que estimamos son ‘locales’ o ‘turistas’

Inequaligram



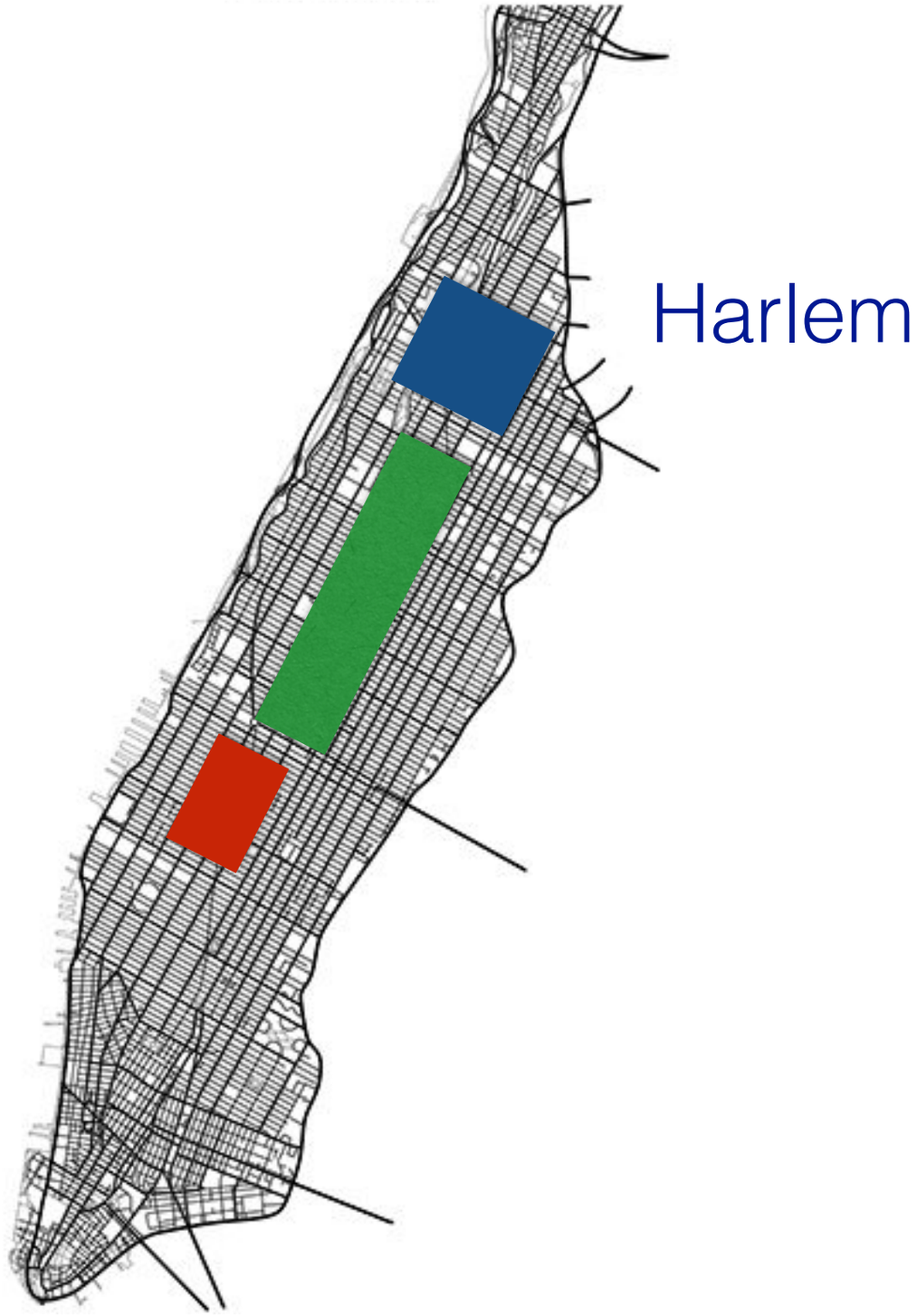
Inequaligram



Inequaligram

Times Square
y Broadway

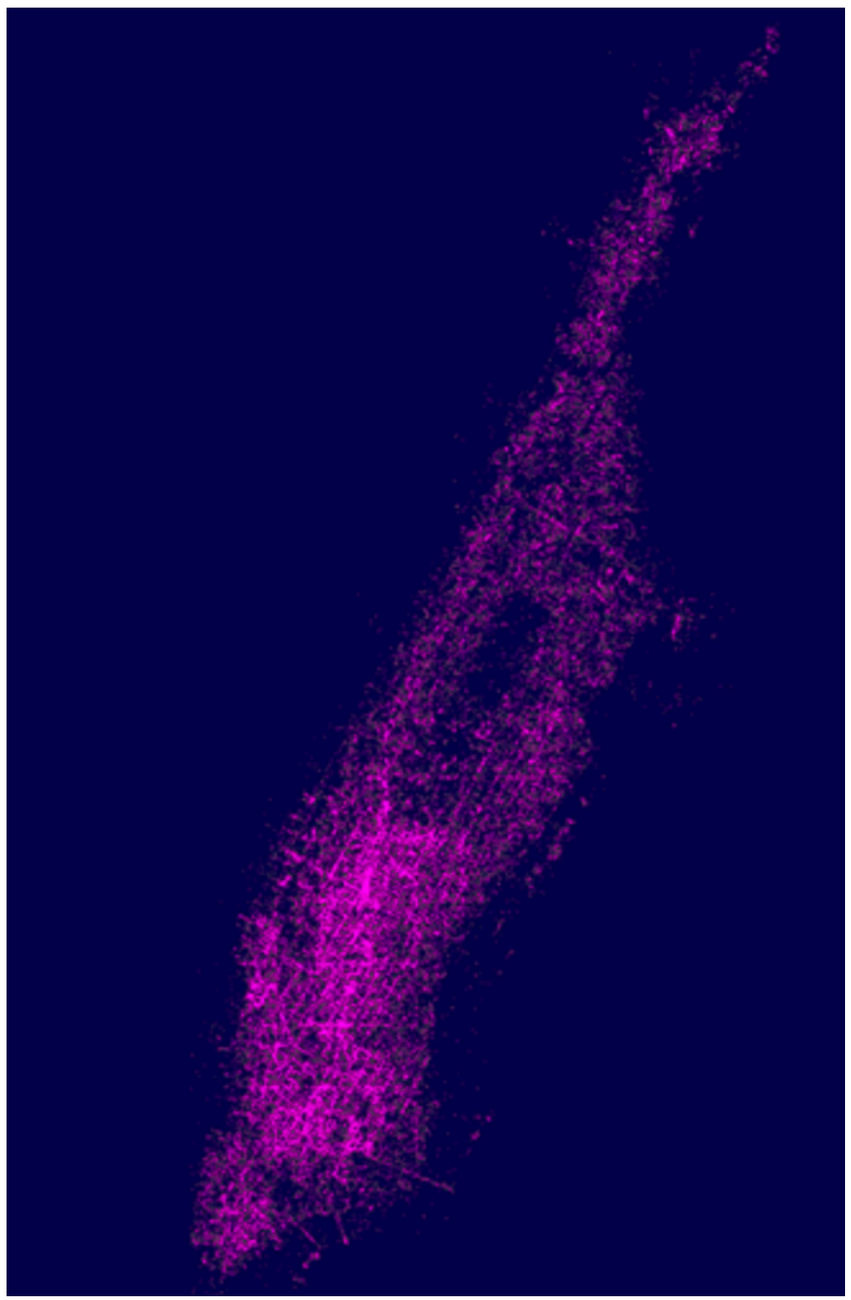
Inequaligram



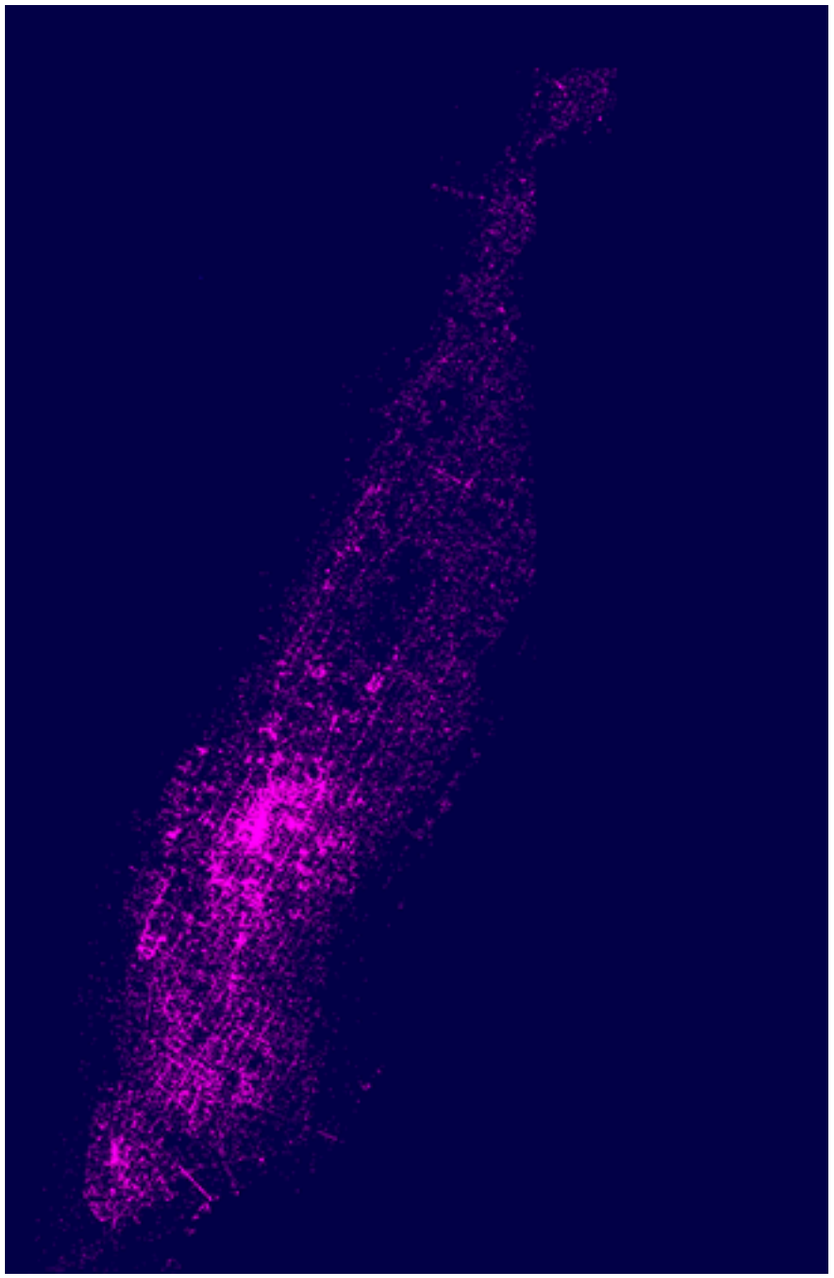
Inequaligram

Locales: distribución 200mil fotos

- concentración de fotos en **midtown y downtown**
- pero hay fotos por toda la isla, casi **no hay zonas invisibles**



Inequaligram

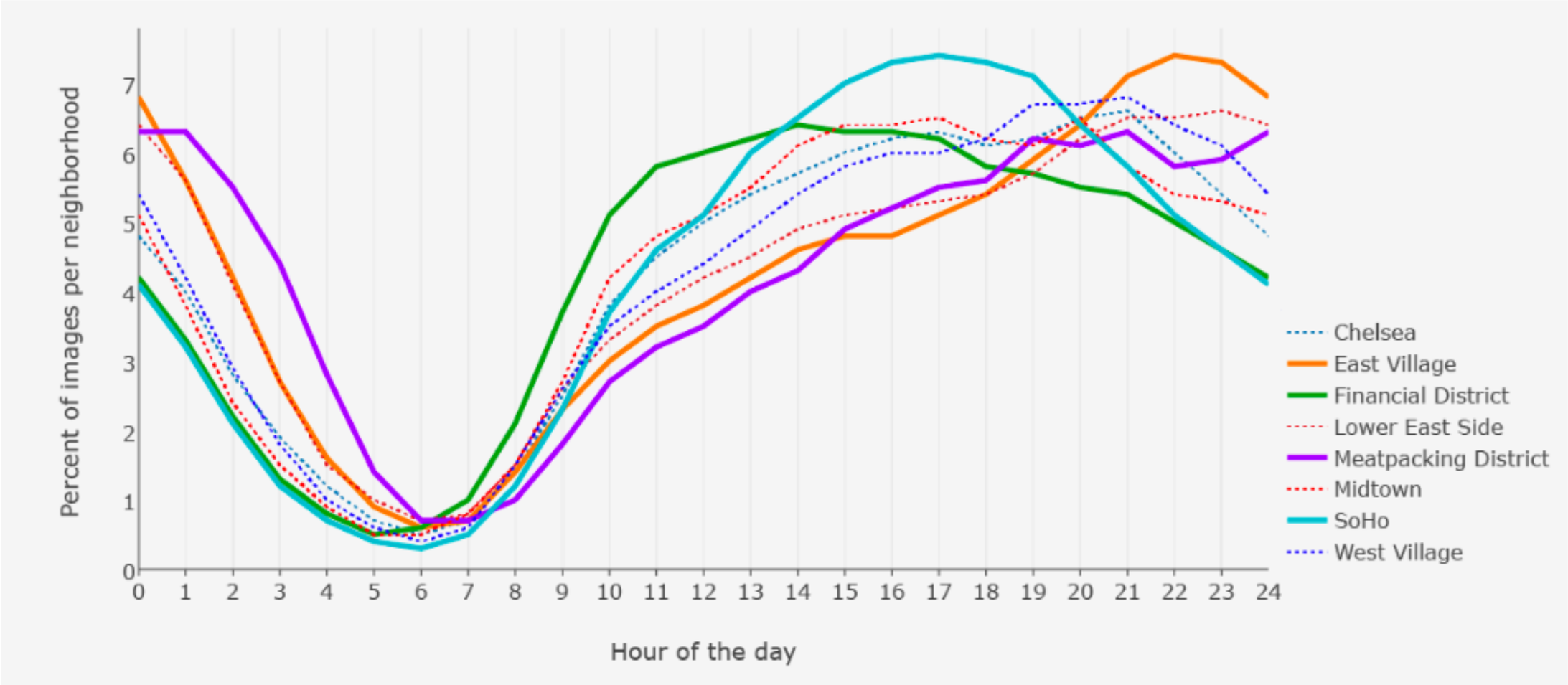


Turistas: distribución 200mil fotos

- concentración de fotos en **midtown**
- pocas imágenes en Harlem

Inequaligram

Distribución de fotos entre locales
por hora y por barrio



Analizando cierre de subte en NY

gothamist

Interactive Maps Give Us A Glimpse Into The Looming L Train Shutdown

BY CHRISTOPHER ROBBINS IN NEWS ON MAR 18, 2016 2:10 PM

Fuente: http://gothamist.com/2016/03/18/l_train_takes_instagram_down.php

Analizando cierre de subte en NY

Analizando cierre de subte en NY

- Alrededor de **50mil personas diarias** toman el subte línea L en alguna de las 3 estaciones del barrio en un día de semana.

- Hay otras líneas de subte que tienen estaciones en el barrio (líneas J, M, Z y G), pero la L es la principal y con buen acceso a Manhattan.

- Entre Brooklyn y Manhattan la línea L pasa por un túnel debajo del agua. **El huracán Sandy (2012) dañó la estructura del túnel y debe ser reparado.** Para ello decidieron **cortar el servicio del L por 18 meses** empezando en enero 2019.

Analizando cierre de subte en NY

With the L Train Closing, Some Shrewd Brooklyn Minds Are Looking Around the J, M, Z Lines

BY LAUREN ELKIES SCHRAM MARCH 8, 2017 11:30 AM

METRO

2019 is the year Williamsburg dies

OPINION

Why the L train closure could be good for Williamsburg

BY REW • NOVEMBER 18, 2016

Politics & Government

East New Yorkers: L Train Closure Will Affect Us, Too

The L Train Shutdown Is a Crisis New York City Can't Afford to Waste

By Henry Grabar

423 112 44

DAILY SHOUTS

THE L TRAIN CLOSURE: HOW WILL IT AFFECT YOU?

N.Y. / REGION

L Train Will Shut Down From Manhattan to Brooklyn in '19 for 18 Months

Analizando cierre de subte en NY

- Juntamos alrededor **500mil fotos subidas a Instagram** desde Williamsburg, Brooklyn en 2014.

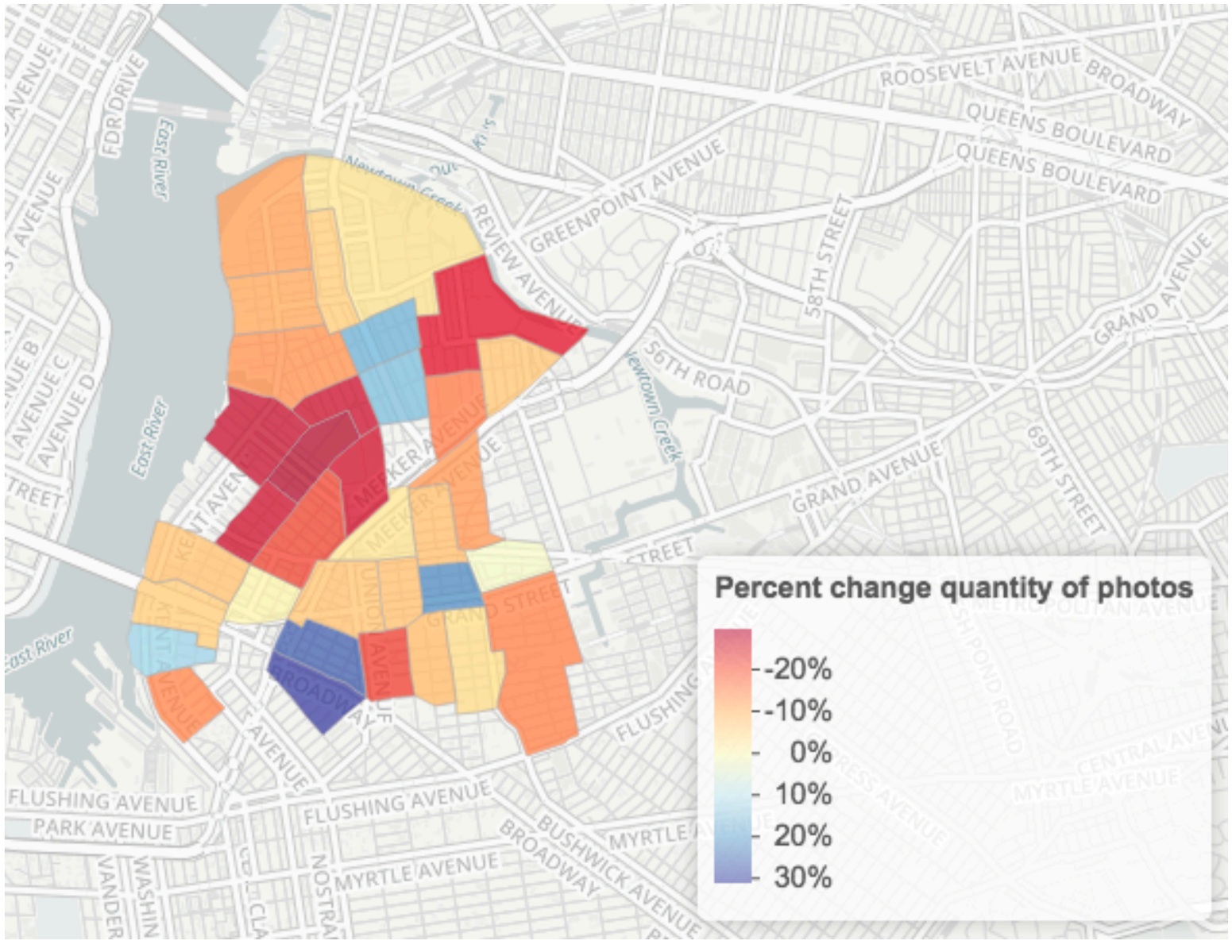
- Agrupamos fotos por zonas censales.

- Calculamos **cantidad de fotos subidas por zona censal** en un **día promedio** con subte L funcionando con normalidad.

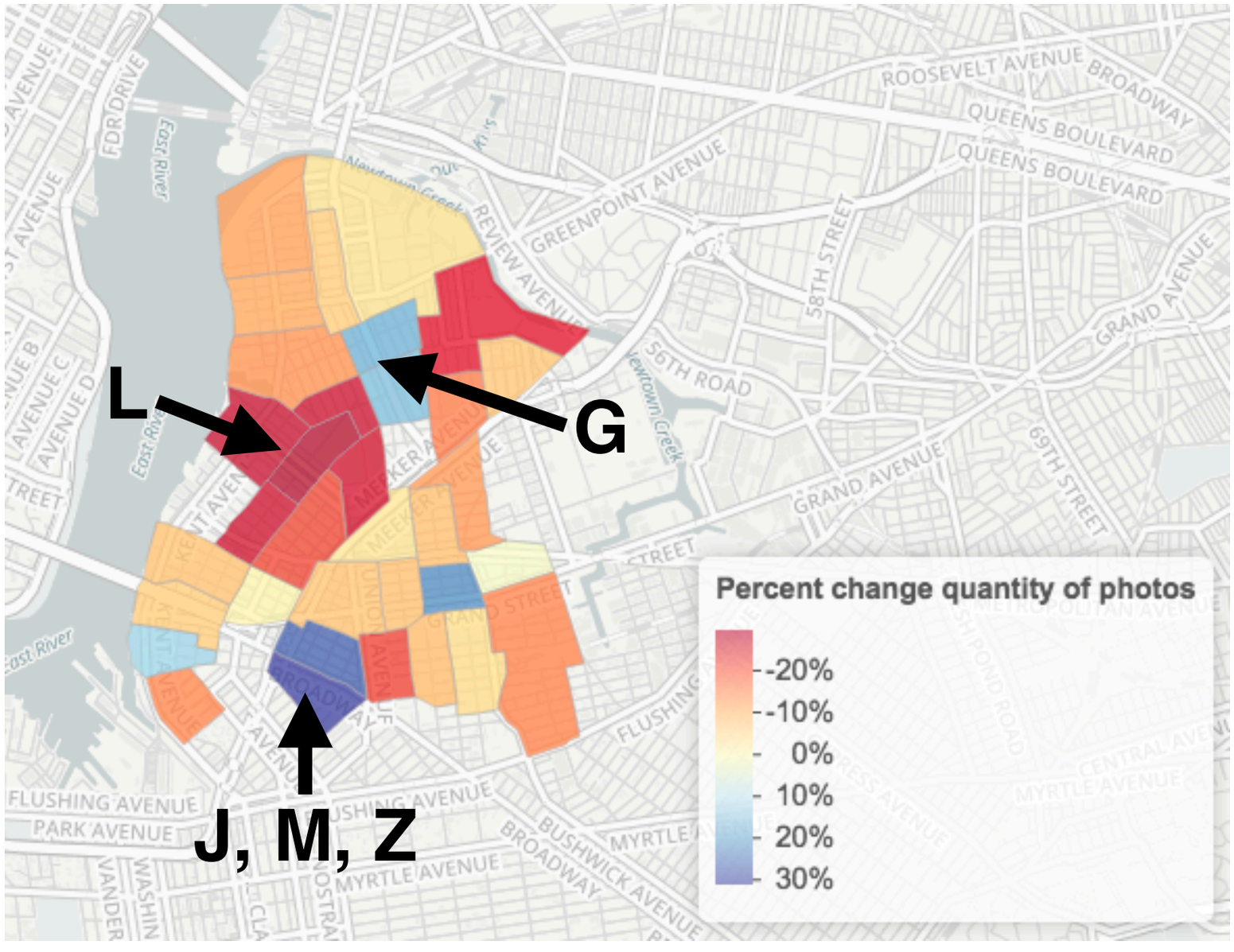
- Calculamos **cantidad de fotos subidas por zona censal** en un día en el cual el **subte L fue interrumpido por mantenimiento.**

- Calculamos la **diferencia en cantidad de fotos** por zona censal **entre el día promedio con el L funcionando normal y el día sin L.**

Analizando cierre de subte en NY

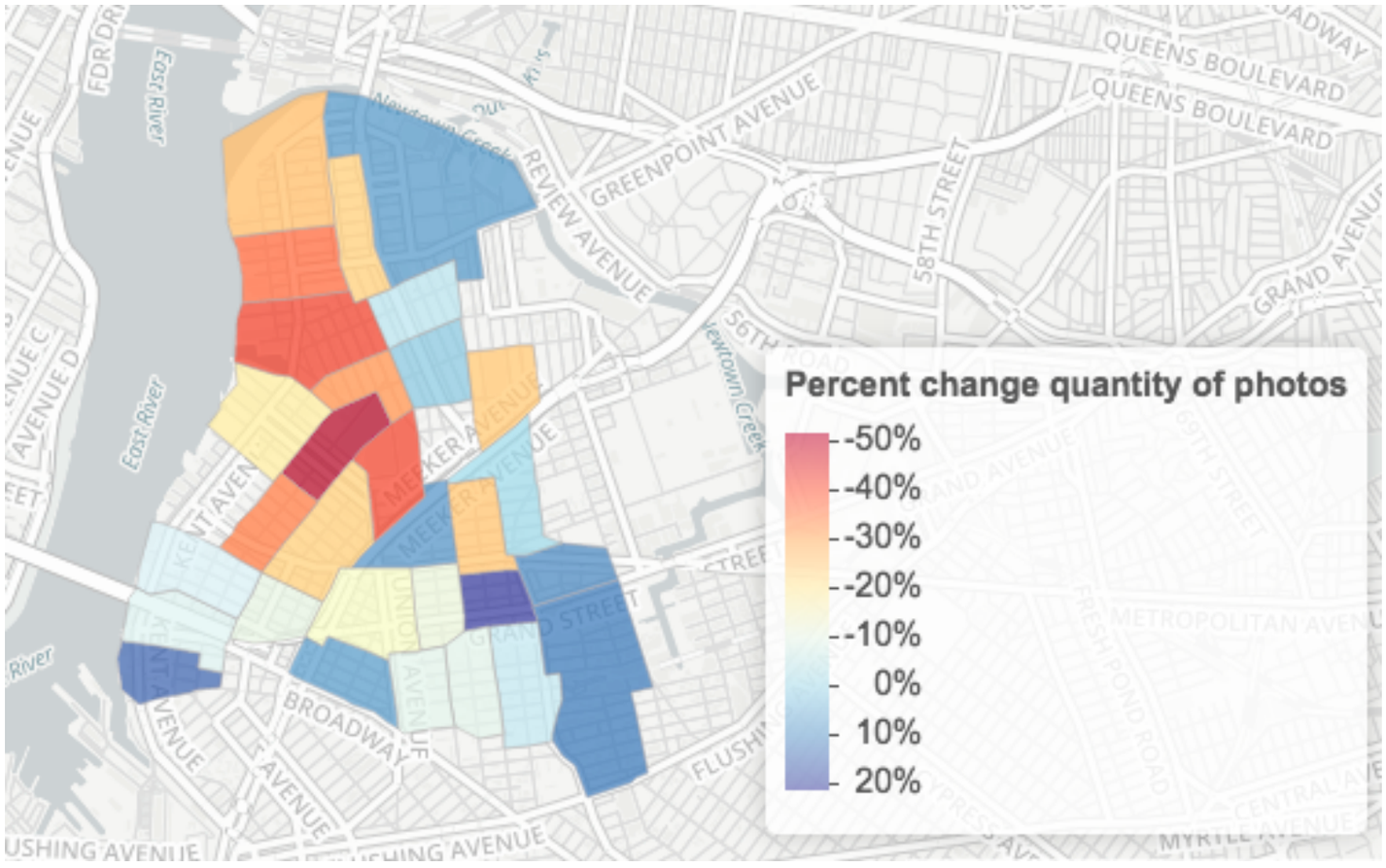


Analizando cierre de subte en NY



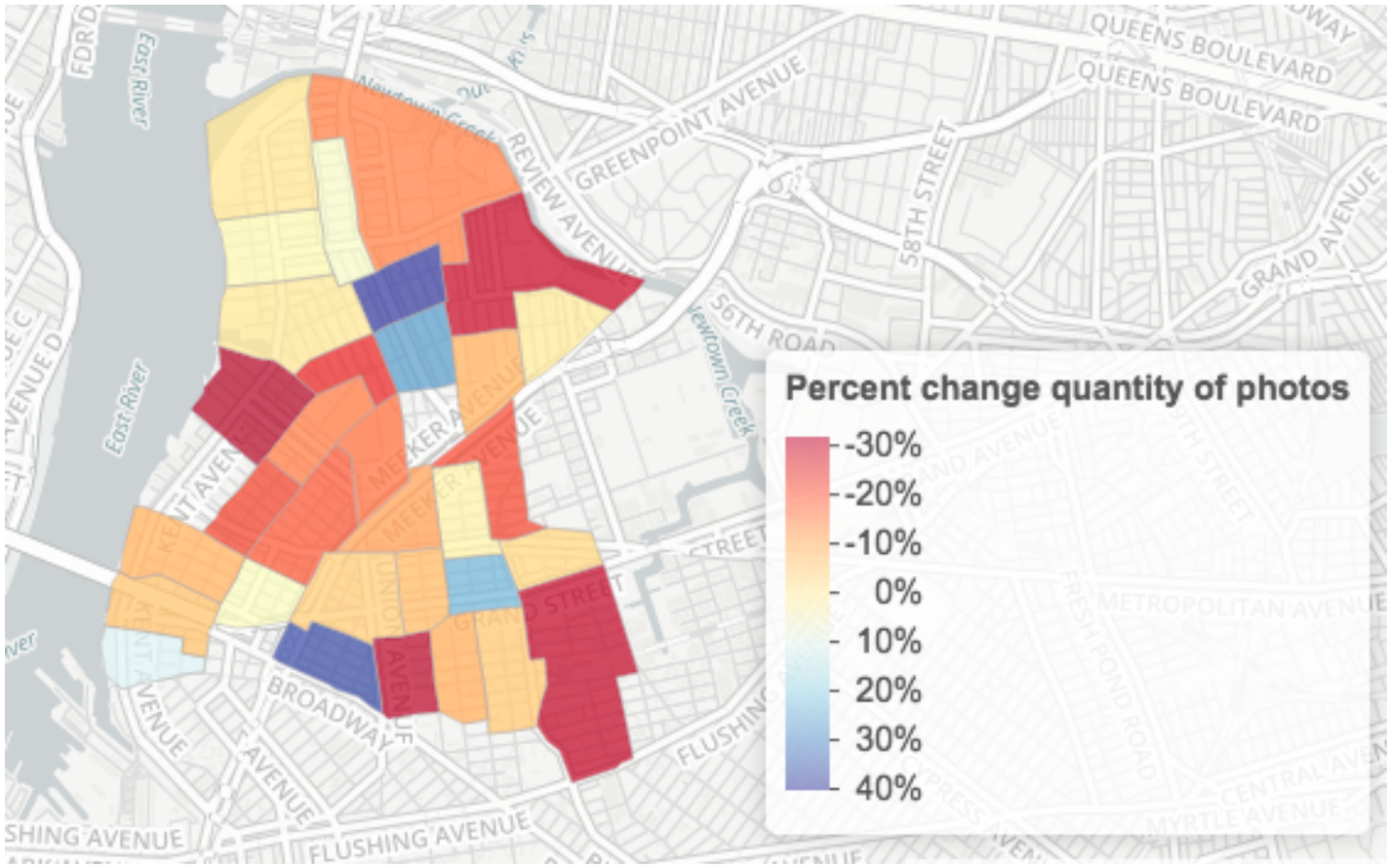
Analizando cierre de subte en NY

Diferencia de día (8am-8pm)

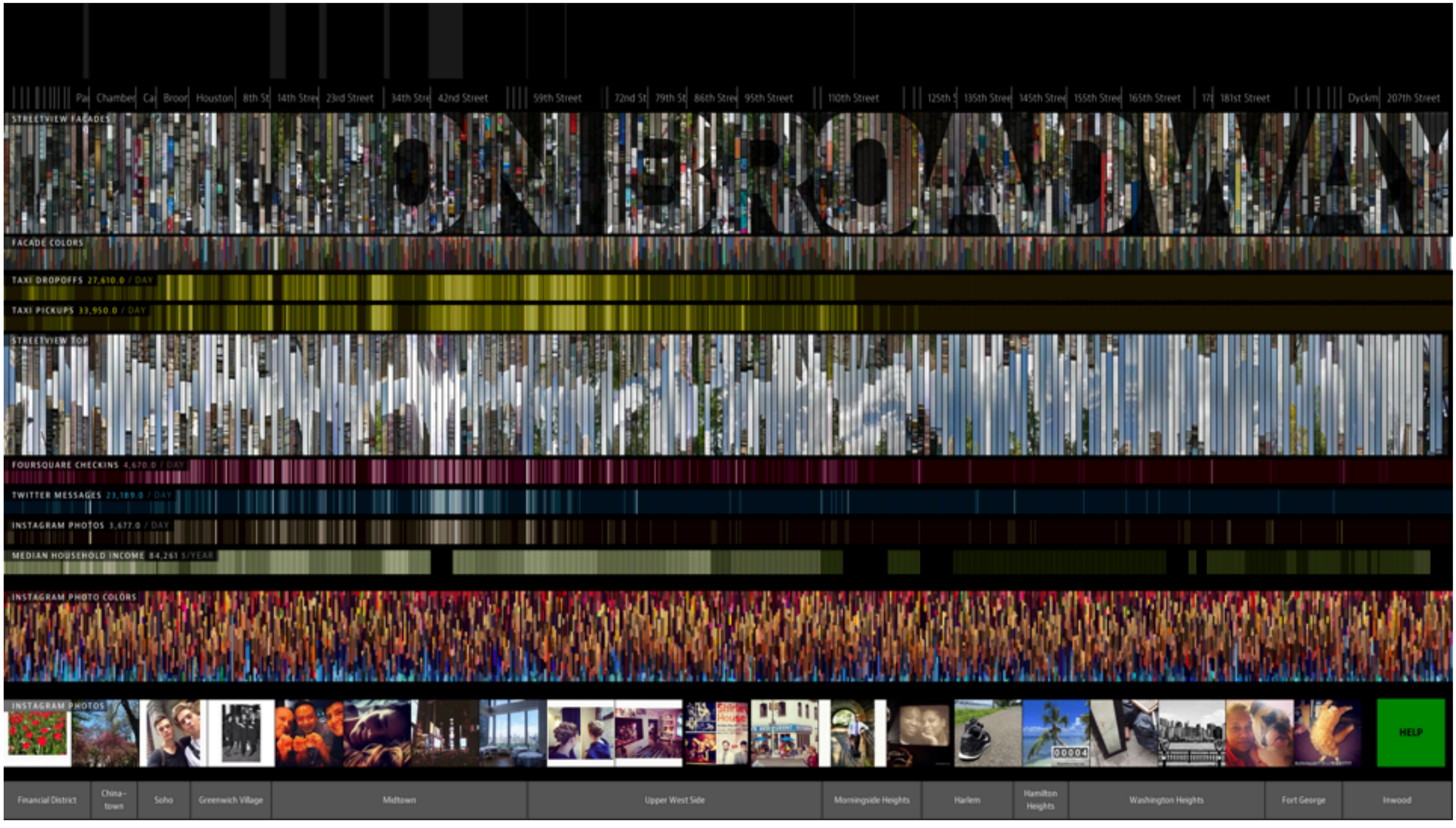


Analizando cierre de subte en NY

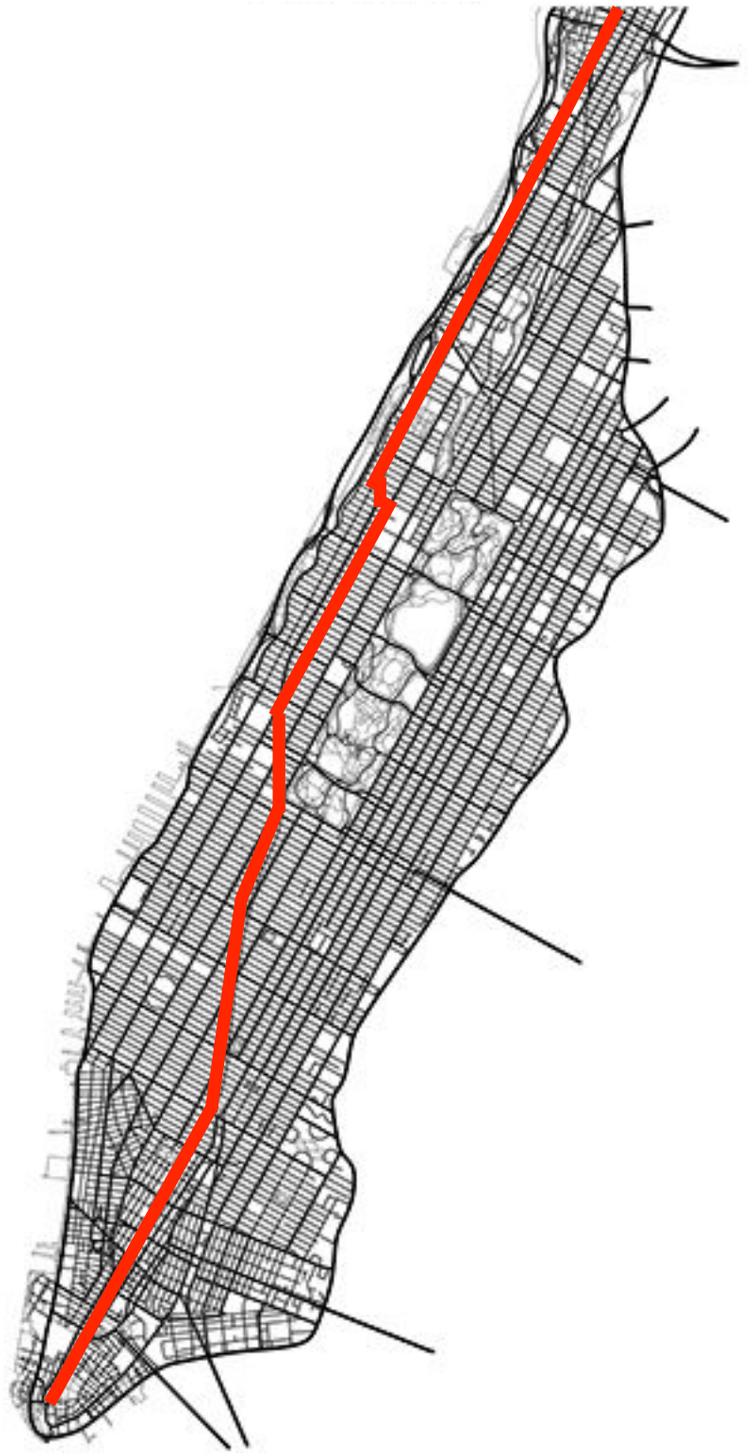
Diferencia de noche (8pm-8am)



On Broadway


Pa
Chambe
Ca
Broor
Houston
8th St
14th Stre
23rd Street
34th Stre
42nd Street
59th Street
72nd St
79th St
86th Stre
95th Street
110th Street
125th S
135th Stre
145th Stre
155th Stre
165th Street
17
181st Street
Dyckm
207th Street
STREETVIEW FACADES
ON BROADWAY
FACADE COLORS
TAXI DROPOFFS 27,610.0 / DAY
TAXI PICKUPS 33,950.0 / DAY
STREETVIEW TOP
FOURSQUARE CHECKINS 4,670.0 / DAY
TWITTER MESSAGES 23,189.0 / DAY
INSTAGRAM PHOTOS 3,677.0 / DAY
MEDIAN HOUSEHOLD INCOME 84,261 $/YEAR
INSTAGRAM PHOTO COLORS
INSTAGRAM PHOTOS
HELP
Financial District
China-town
Soho
Greenwich Village
Midtown
Upper West Side
Morningside Heights
Harlem
Hamilton Heights
Washington Heights
Fort George
Inwood

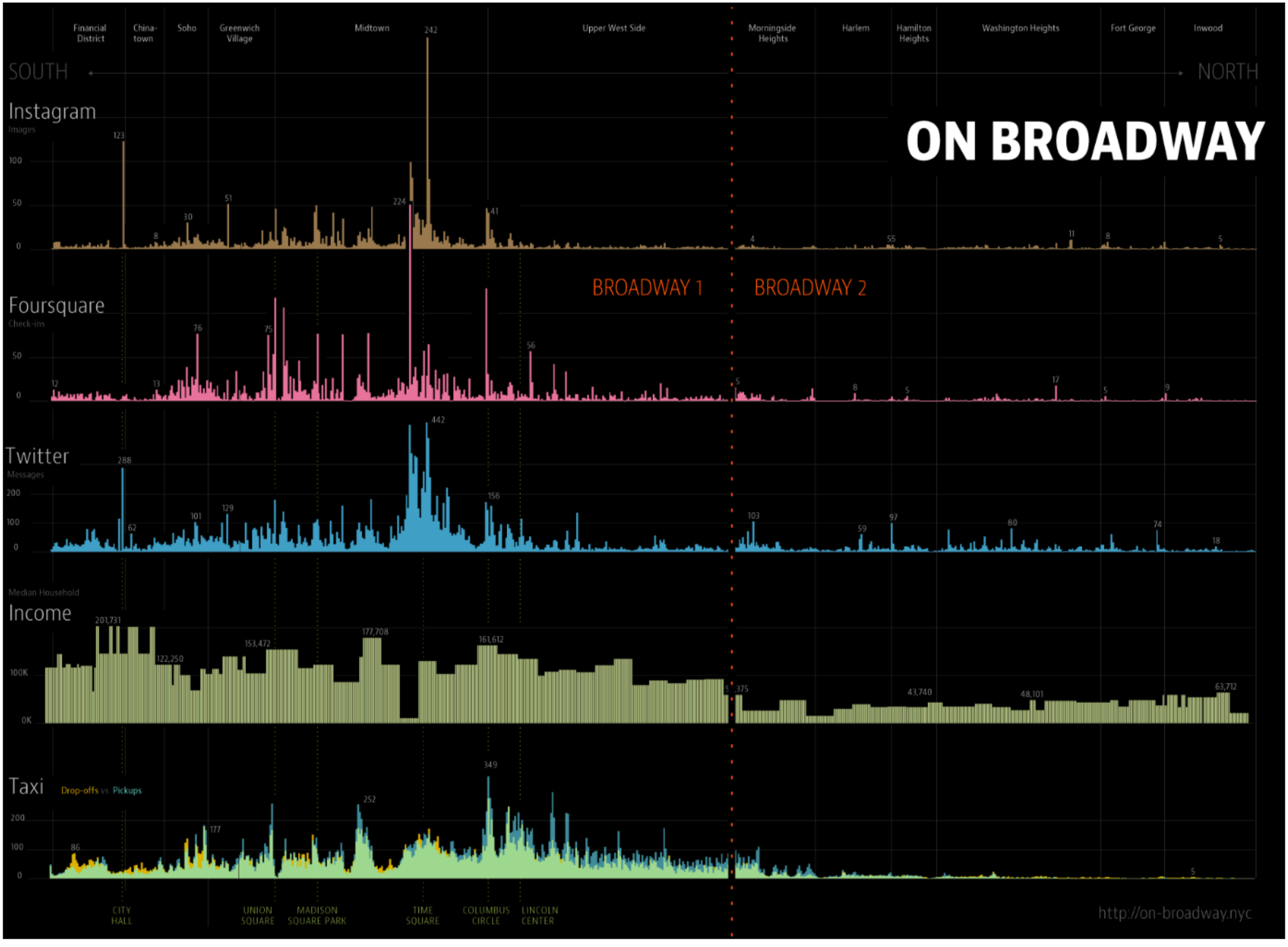
On Broadway



On Broadway

La ciudad del siglo XXI nos habla en datos:

- **Instagram** (10.5millones, 650mil en Broadway)
- **Tweets** (14millones, 1millón en Broadway)
- **Foursquare** (8millones de checkins en Broadway)
- **Taxis** (140millones, 12millones en Broadway)
- **Ingreso y alquiler medio** (datos de censo)


Financial District
China-town
Soho
Greenwich Village
Midtown
Upper West Side
Morningside Heights
Harlem
Hamilton Heights
Washington Heights
Fort George
Inwood
SOUTH
NORTH
ON BROADWAY
Instagram
Images
123
100
50
0
8
30
51
224
242
41
4
55
11
8
5
BROADWAY 1
BROADWAY 2
Foursquare
Check-ins
50
0
12
13
76
75
56
5
8
5
17
5
9
Twitter
Messages
288
200
100
0
62
101
129
442
156
103
59
97
80
74
18
Median Household
Income
201,731
122,250
153,472
177,708
161,612
100K
0K
,375
43,740
48,101
63,712
Taxi
Drop-offs vs. Pickups
200
100
0
86
177
252
349
5
CITY HALL
UNION SQUARE
MADISON SQUARE PARK
TIME SQUARE
COLUMBUS CIRCLE
LINCOLN CENTER
http://on-broadway.nyc

On Broadway

Video: https://vimeo.com/118247767

Más información sobre proyectos

Inequaligram:

- Indaco A. y Manovich, L. *Urban Social Media Inequality: Definition, Measurements and Applications,* Urban Studies and Practices Journal, Volume 87, Issue 1, Feb, 2017.

- http://inequaligram.net/

L train:

- http://rpubs.com/aindaco/155270

On Broadway:

- http://www.on-broadway.nyc/