

Geographic Data Science - Lecture II (New) Spatial Data

Dani Arribas-Bel

"Yesterday"

- Introduced the (geo-)data revolution
 - What is it?
 - Why now?
- The *need* of (geo-)data science to make sense of it all

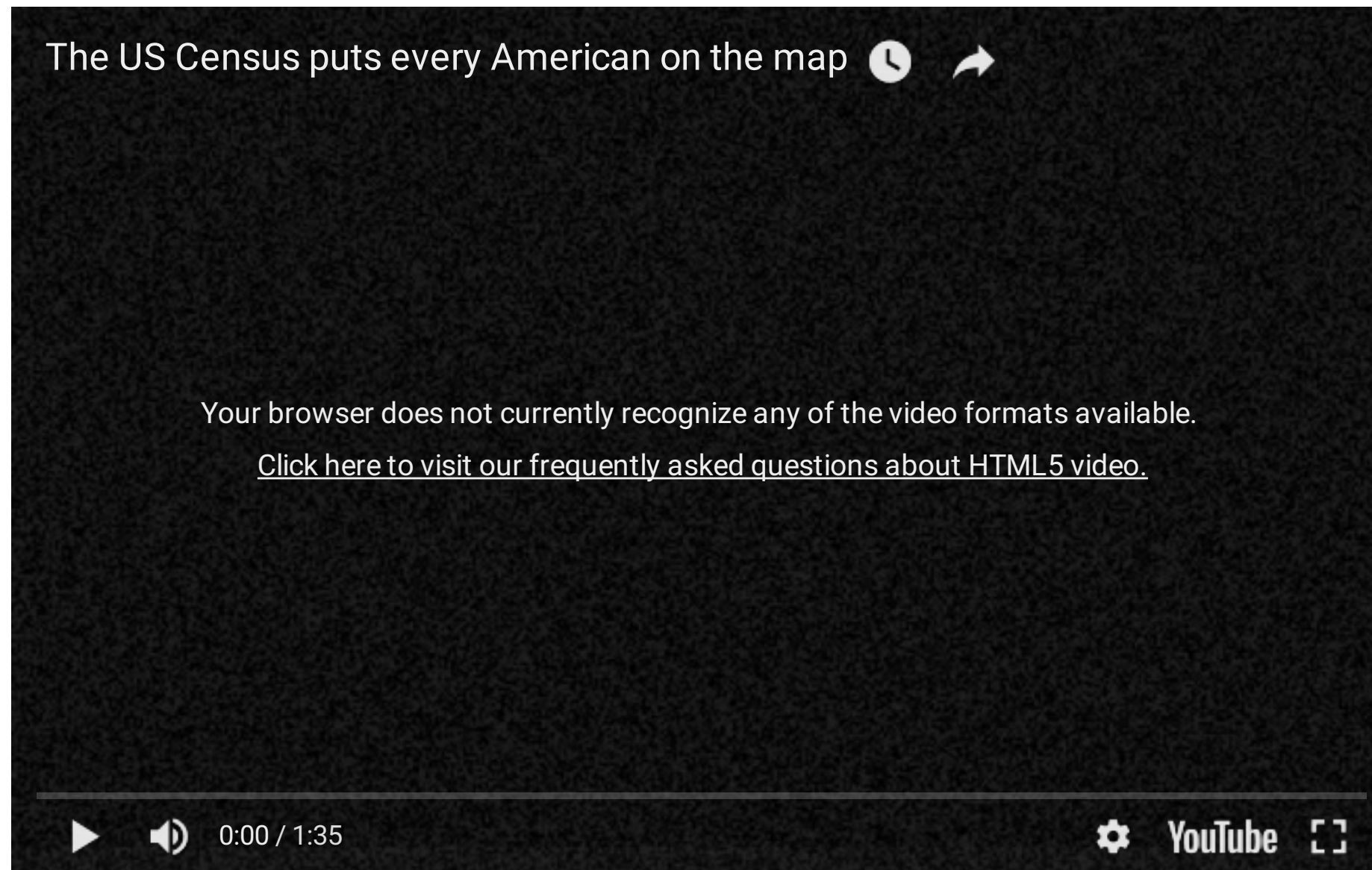
Today

- Traditional data: refresher
- New sources of spatial data
- Opportunities & Challenges

Good old spatial data

Good old spatial data

[[source](#)]



Good old spatial data (+)

Traditionally, datasets used in the (social) sciences are:

- Collected for the purpose --> **carefully** designed
- **Detailed** in information ("*...rich profiles and portraits of the country...*")
- **High quality**

Good old spatial data (-)

But also:

- Massive enterprises ("*...every single person...*") --> costly
- But coarse in resolution (to preserve privacy they need to be aggregated)
- Slow: the more detailed, the less frequent they are available

Examples

- Decennial census (and census geographies)
- Longitudinal surveys
- Customly collected surveys, interviews, etc.
- Economic indicators
- ...

New sources of (spatial) data

New sources of (spatial) data

Tied into the (geo-)data revolution, new sources are appearing that are:

- **ACCIDENTAL** --> created for different purposes but available for analysis as a side effect
- Very diverse in nature, resolution, and detail but, potentially, much more **detailed** in both space and time
- Quality also varies greatly

Different ways to categorise them...

Lazer & Radford (2017)

- Digital Life: digital actions (Twitter, Facebook, Wikipedia...)
- Digital traces: record of digital actions (CDRs, metadata...)
- Digitalised life: nonintrinsically digital life in digital form (Government records, web...)

Arribas-Bel (2014)

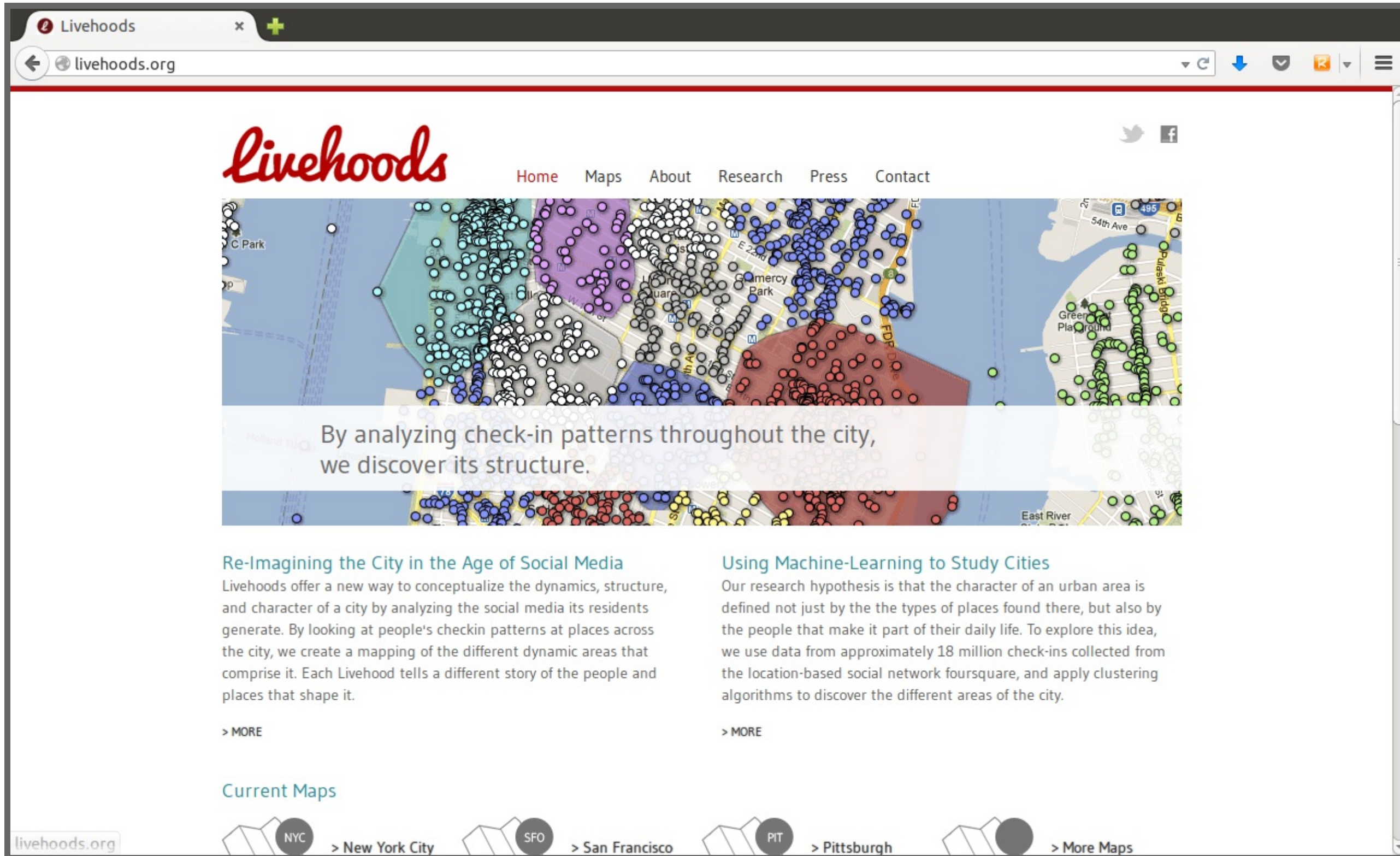
Three levels, based on how they originate:

- [Bottom up] "Citizens as sensors"
- [Intermediate] Digital businesses/businesses going digital
- [Top down] Open Government Data

Citizens as sensors

- Technology has allowed widespread adoption of sensors (bands, smartphones, tablets...)
- (Almost) every aspect of human life is subject to leave a digital trace that can be collected, stored and analyzed
- Individuals become content/data creators (sensors, *Goodchild, 2007*)
- *Why relevant for geographers?* --> Most of it (80%?) has some form of spatial dimension

Example: Livehoods



The screenshot shows the Livehoods website interface. At the top, there is a browser window with the address bar showing "livehoods.org". The website header features the "livehoods" logo in red script, followed by a navigation menu with links for "Home", "Maps", "About", "Research", "Press", and "Contact". Social media icons for Twitter and Facebook are also present. The main content area is dominated by a map of New York City, where various neighborhoods are highlighted in different colors (purple, blue, red, green) and populated with small circular markers representing check-in data. Below the map, a white banner contains the text: "By analyzing check-in patterns throughout the city, we discover its structure." Underneath this banner, there are two columns of text. The left column is titled "Re-Imagining the City in the Age of Social Media" and describes how Livehoods use social media check-in data to map city dynamics. The right column is titled "Using Machine-Learning to Study Cities" and explains the research hypothesis and the use of machine learning on Foursquare data. Both columns end with a "> MORE" link. At the bottom of the page, there is a "Current Maps" section with icons and links for "NYC > New York City", "SFO > San Francisco", "PIT > Pittsburgh", and "> More Maps". The "livehoods.org" logo is visible in the bottom left corner.

livehoods

Home Maps About Research Press Contact

By analyzing check-in patterns throughout the city, we discover its structure.

Re-Imagining the City in the Age of Social Media
Livehoods offer a new way to conceptualize the dynamics, structure, and character of a city by analyzing the social media its residents generate. By looking at people's checkin patterns at places across the city, we create a mapping of the different dynamic areas that comprise it. Each Livehood tells a different story of the people and places that shape it.
> MORE

Using Machine-Learning to Study Cities
Our research hypothesis is that the character of an urban area is defined not just by the the types of places found there, but also by the people that make it part of their daily life. To explore this idea, we use data from approximately 18 million check-ins collected from the location-based social network foursquare, and apply clustering algorithms to discover the different areas of the city.
> MORE

Current Maps

livehoods.org

NYC > New York City SFO > San Francisco PIT > Pittsburgh > More Maps

Businesses moving online

- Many of the elements and parts of business activities have been computerized in the last decades
- This implies, without any change in the final product or activity per se, a lot more digital data is "available" about their operations
- In addition, entirely new business activities have been created based on the new technologies ("internet natives")
- Much of these data can help researchers better understand how cities work

Example: Walkscore

San Francisco Apart... x

https://www.walkscore.com/CA/San_Francisco

Walk Score **84** San Francisco is Very Walkable
Most errands can be accomplished on foot.

Walk Score Map

Sutro Baths

Presidio San Francisco

San Francisco is the 2nd most walkable large city in the US with 805,235 residents.

San Francisco has excellent public transportation and is very bikeable.

Find apartments in San Francisco's most walkable neighborhoods: [Chinatown](#), [Financial District](#) and [Downtown](#).

[San Francisco Apartments for Rent](#) [San Francisco Homes for Sale](#)

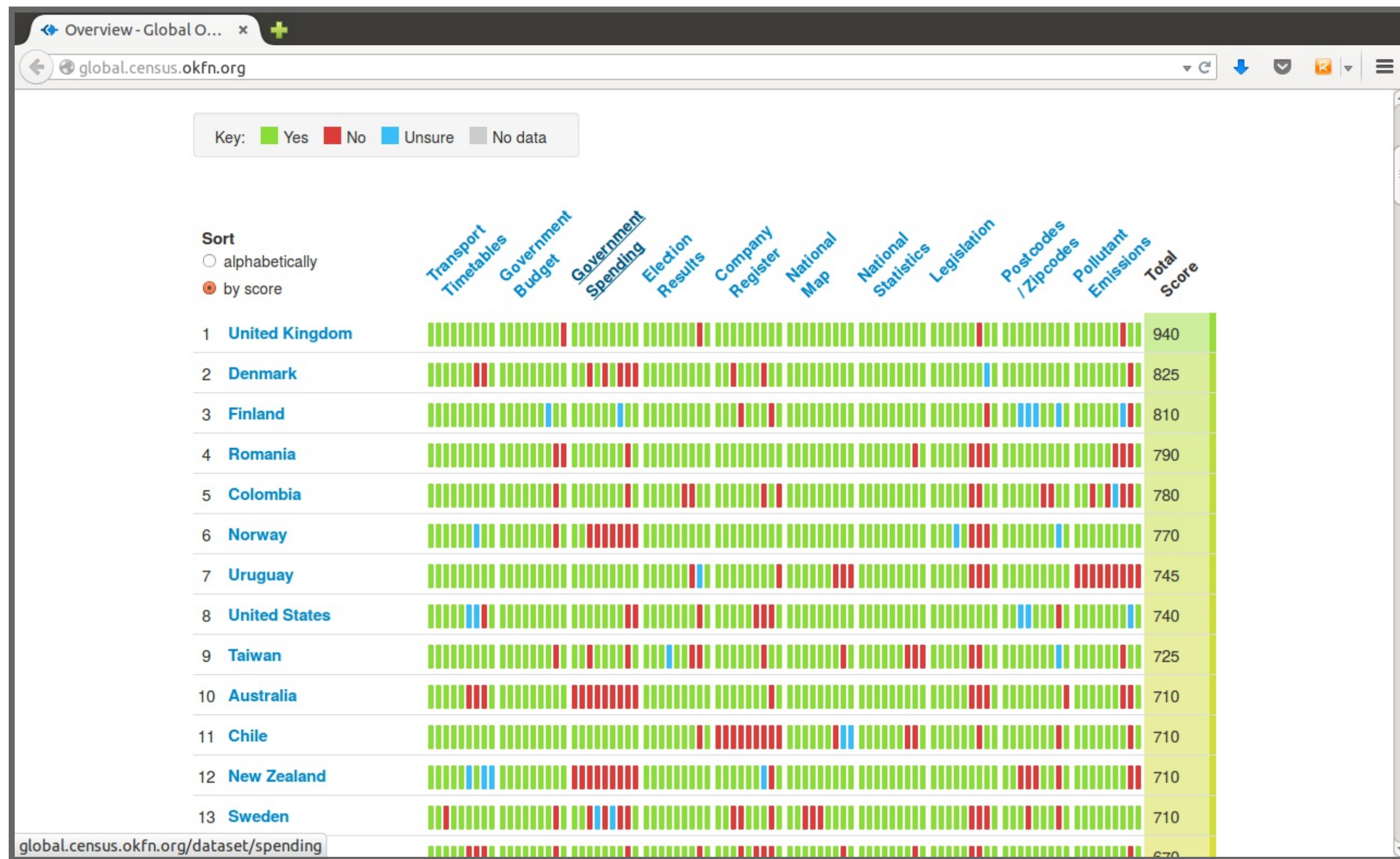
View all [San Francisco apartments](#) on a map. The average rent is \$3,750 and the average home price is \$1,099,999. ?

Open data for open governments

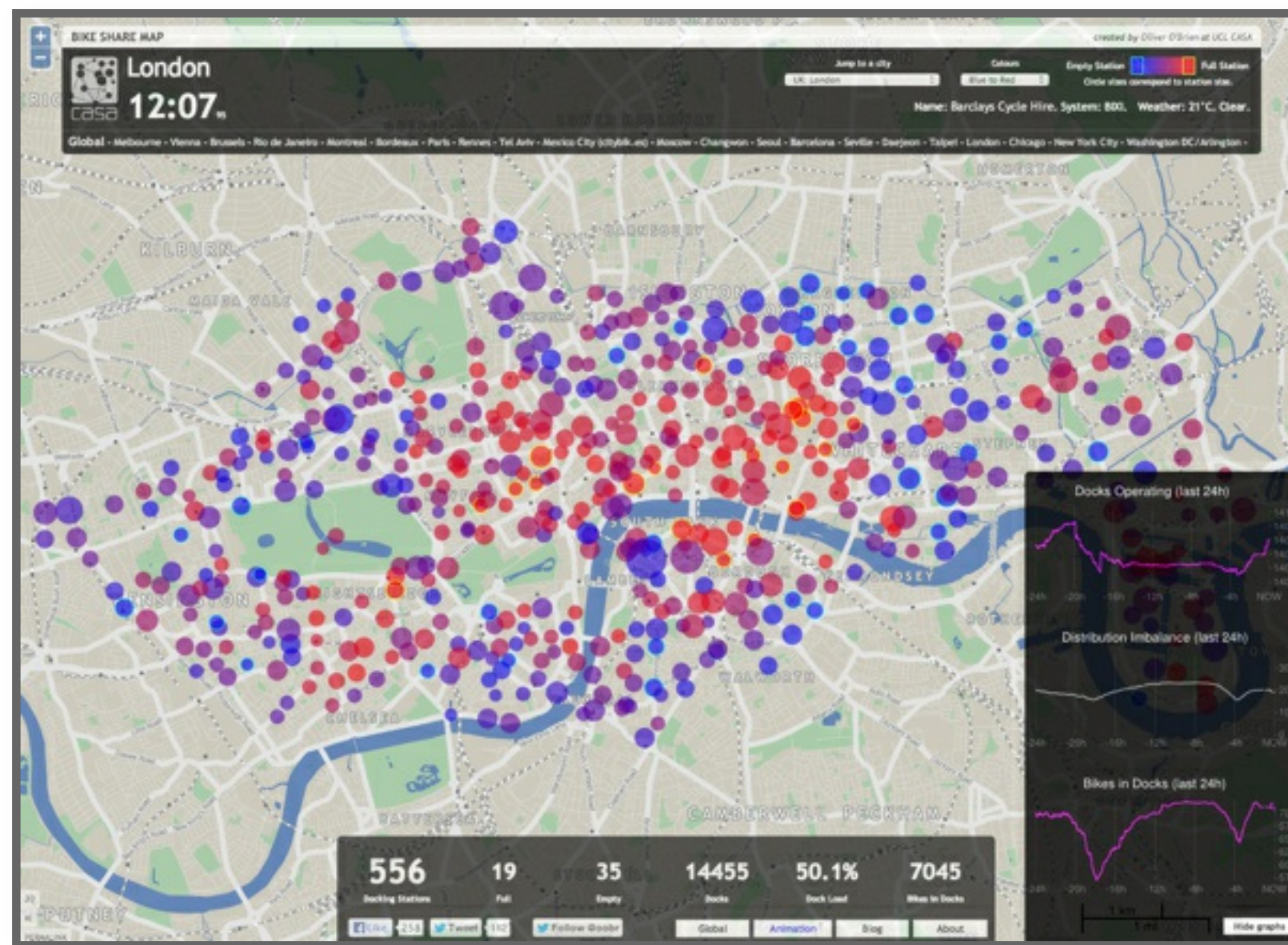
Government institutions release (part of) their internal data in open format. Motivations ([Shadbolt, 2010](#)):

- Transparency and accountability
- Economic and social value
- Public service improvement
- Creation of new industries and jobs

Global Open Data Index'14



Example: BikeShare Map



Class Quiz

Class Quiz

In pairs, 2 minutes to discuss the origin of the following sources of (geo-)data:

- Geo-referenced tweets --> Bottom-up
- Land-registry house transaction values --> Open Government
- Google maps restaurant listing --> Digital businesses
- ONS Deprivation Indices --> Traditional (not accidental!)
- Liverpool bikeshare service station status --> Open Government Data

Opportunities & Challenges

Opportunities

From Lazer & Radford (2017):

- Massive, passive
- Nowcasting
- Data on social systems
- Natural and field experiments ("always-on" observatory of human behaviour)
- Making big data small

Challenges

- Bias
- Technical barriers to access
- The need of new methods

Bias

- Traditional data meet some quality standards (representativity, accuracy...)
- Because they're *accidental*, new data sources might not
- Researchers need to have extra care and put more thought into what conclusions they can reach from analyses with new sources of data
- In some cases, bias can run in favour of researchers, but this should never be taken for granted

Technical barriers to access

- Much of these data are available
- However, their accidental nature makes them not be *directly* available
- Usually, a **different set of skills** is required to tap into their power
 - Basic programming
 - Computing literacy (understanding of the internet, APIs, databases...)
 - Software savvy-ness (a.k.a. "go beyond Word and Excel")

(New) Methods

The nature of these data is not exactly the same as that of more traditional datasets. For example:

- Spatial aggregation: Polygons Vs. Points
- Temporal aggregation(frequency): Decadal Vs. Real-time

Some of this does not "play well" with techniques employed traditionally to analyze data in Geography.

(New) Methods



[source]

(New) Methods

To be able to extract as much insight as possible from these new sources of data --> *borrow* techniques from other disciplines, or even *create* new ones

Examples:

- Visualization
- Machine learning

But also others like bayesian inference, network science...

New + Old

Traditional data:

- High quality, detailed, and reliable
- Costly, coarse, and slow

Accidental data:

- Cheap, fine-grained, and fast
- Less reliable, harder to access, and potentially uninteresting

--> 1 + 1 > 2



Geographic Data Science'15 – Lecture 1 by Dani Arribas-Bel is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.