

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

- Spatial data: (quick) refresher
- New sources of (spatial) data
- Challenges
- (Cool) examples

(Good old) spatial data

(Good old) spatial data

- Types
- Characteristics (+ and -)

New sources of (spatial) data

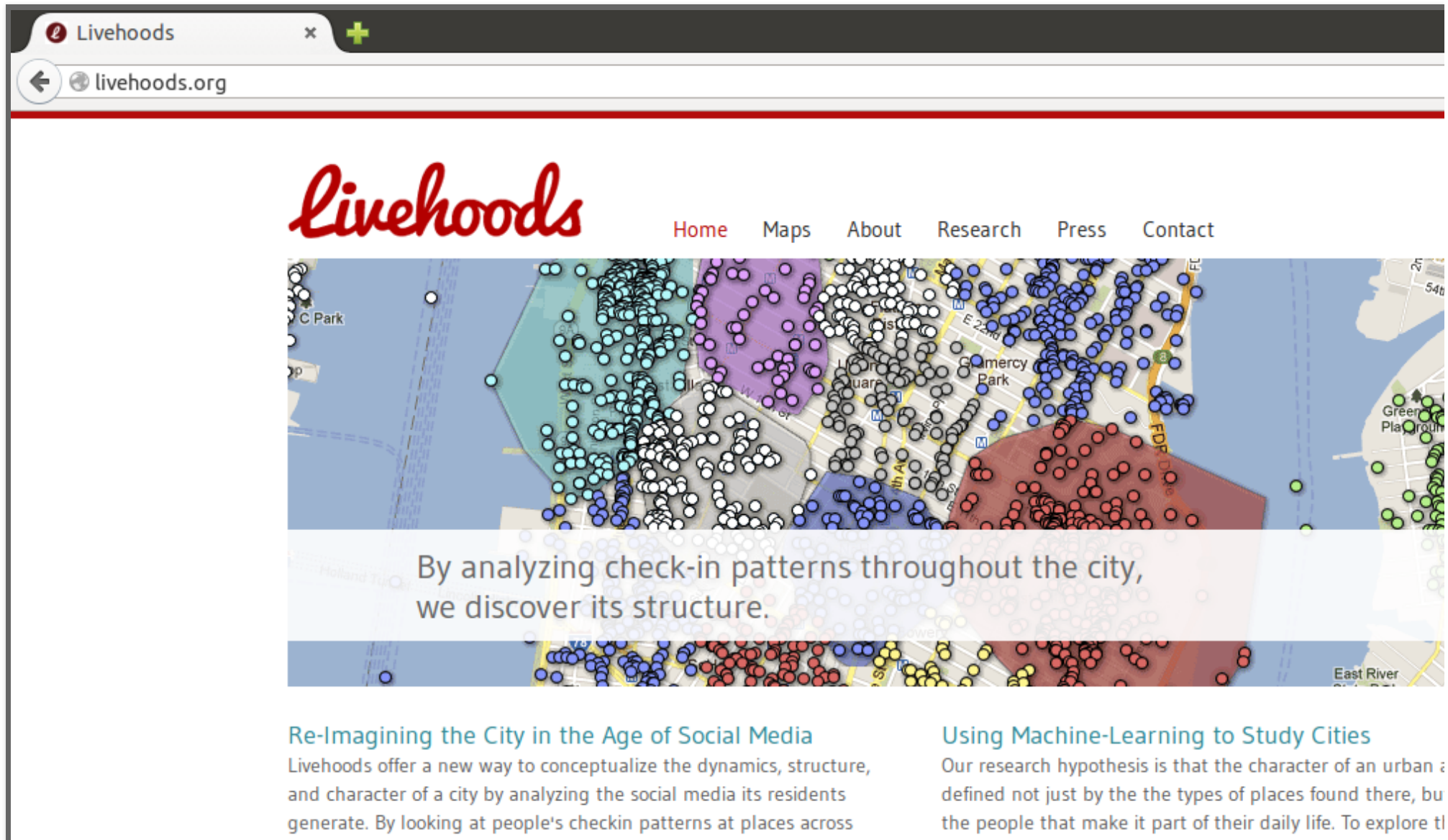
New sources of (spatial) data

- Tied into the (geo-)data revolution
- Multi-type, but falling into same categories (polygons, points, surfaces...)
- Accidental --> very different nature
- Levels at which they are originated:
 - [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 a web browser window with the address bar displaying "livehoods.org". The website features a red "Livehoods" logo and a navigation menu with links to Home, Maps, About, Research, Press, and Contact. Below the navigation is a large map of New York City, specifically the area around the East River and Lower East Side. The map is overlaid with numerous small, colored circles representing check-in data points. These points are clustered into several distinct regions, each highlighted with a different color: a large cyan cluster in the center, a purple cluster to the west, a blue cluster to the east, and a red cluster to the south. A semi-transparent white banner is overlaid on the map with the text: "By analyzing check-in patterns throughout the city, we discover its structure." Below the map, 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 uses social media check-in data to understand city dynamics. The right column is titled "Using Machine-Learning to Study Cities" and discusses the research hypothesis that urban character is defined by the types of places and the people that use them.

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

Using Machine-Learning to Study Cities

Our research hypothesis is that the character of an urban is defined not just by the the types of places found there, but the people that make it part of their daily life. To explore tl

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

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



> New York City



> San Francisco



> Pittsburgh



> More Maps

Business 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 this 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

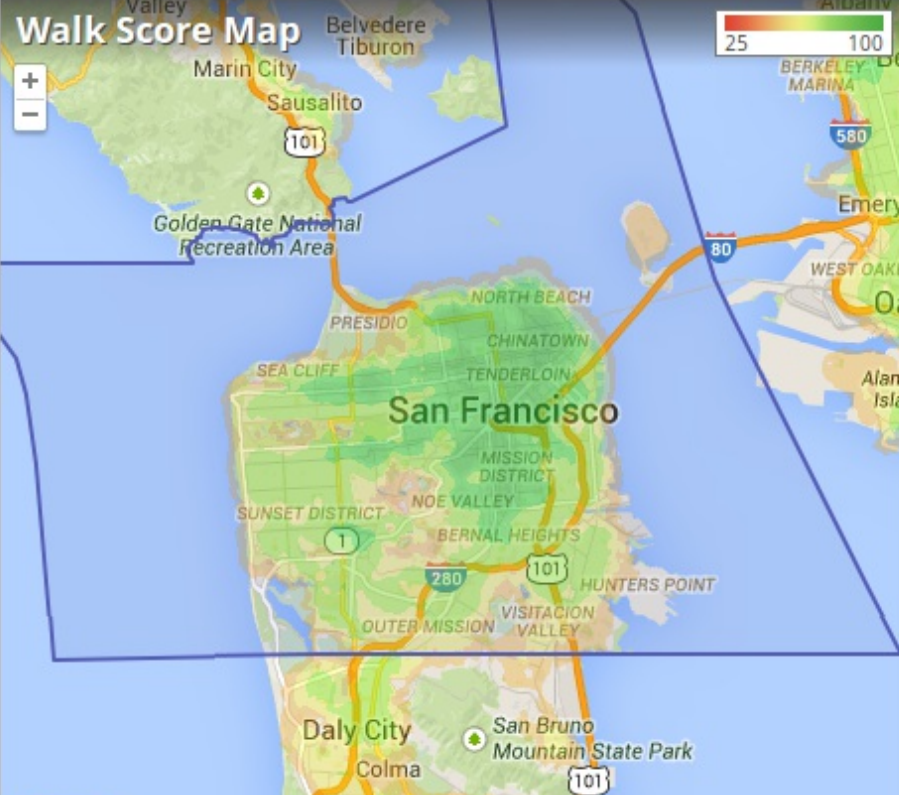
San Francisco is Very Walkable

Most errands can be accomplished on foot.


Walk Score Map

+


-



25 100



Sutro Baths



Presidio San Francisco

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

San Francisco has excellent public transportation bikeable.

Find apartments in San Francisco's most walkable

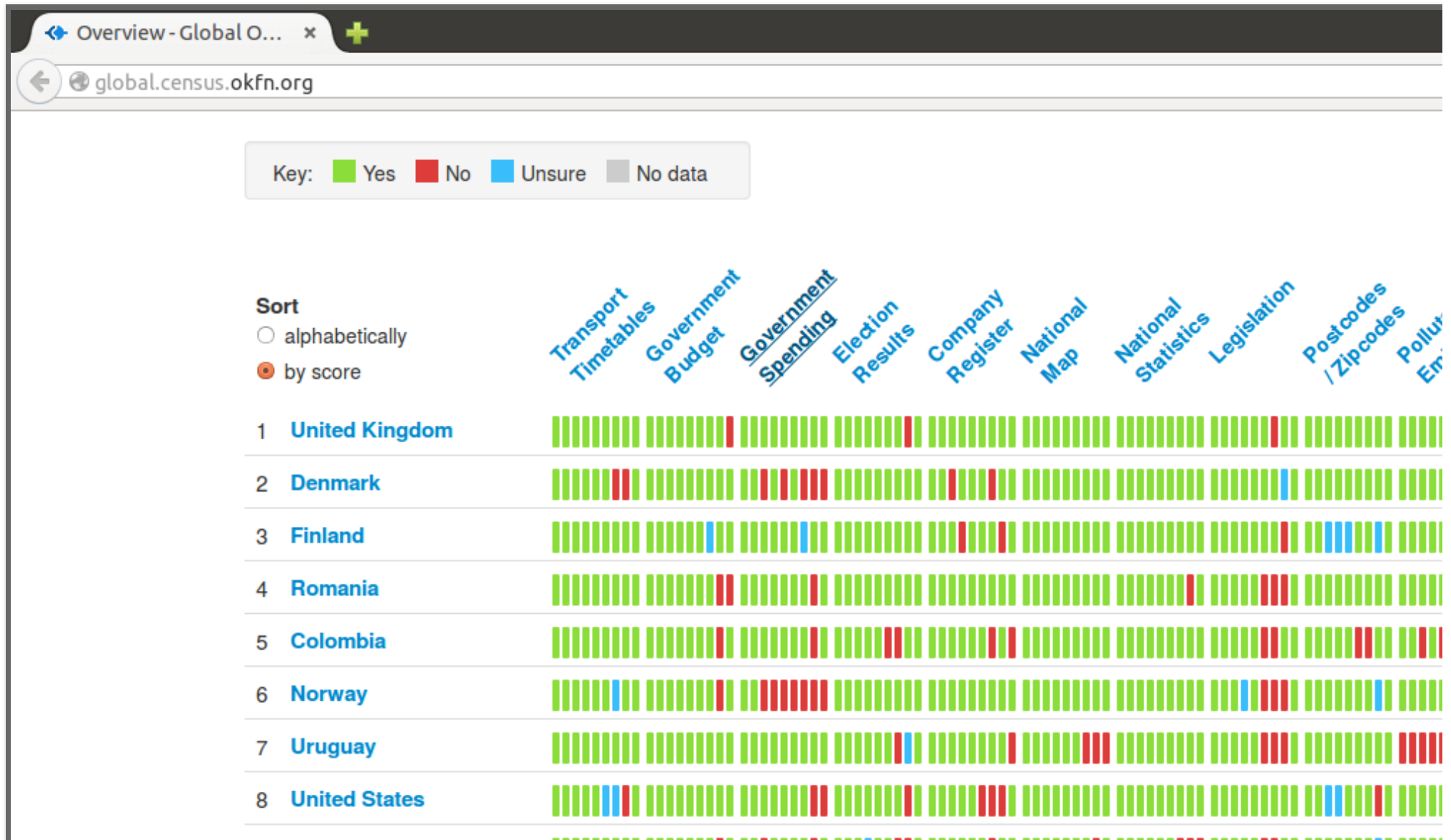
The screenshot shows a Google Maps interface. At the top, the Google logo is on the left, and a map of San Francisco is in the center. To the right of the map, the text "neighborhoods: Chinatown, Financial District and" is visible. Below the map, the text "United States > California > San Francisco" is displayed. In the center, there are two buttons: "San Francisco Apartments for Rent" and "San Francisco Homes for Sa". At the bottom, a text line reads: "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 2014



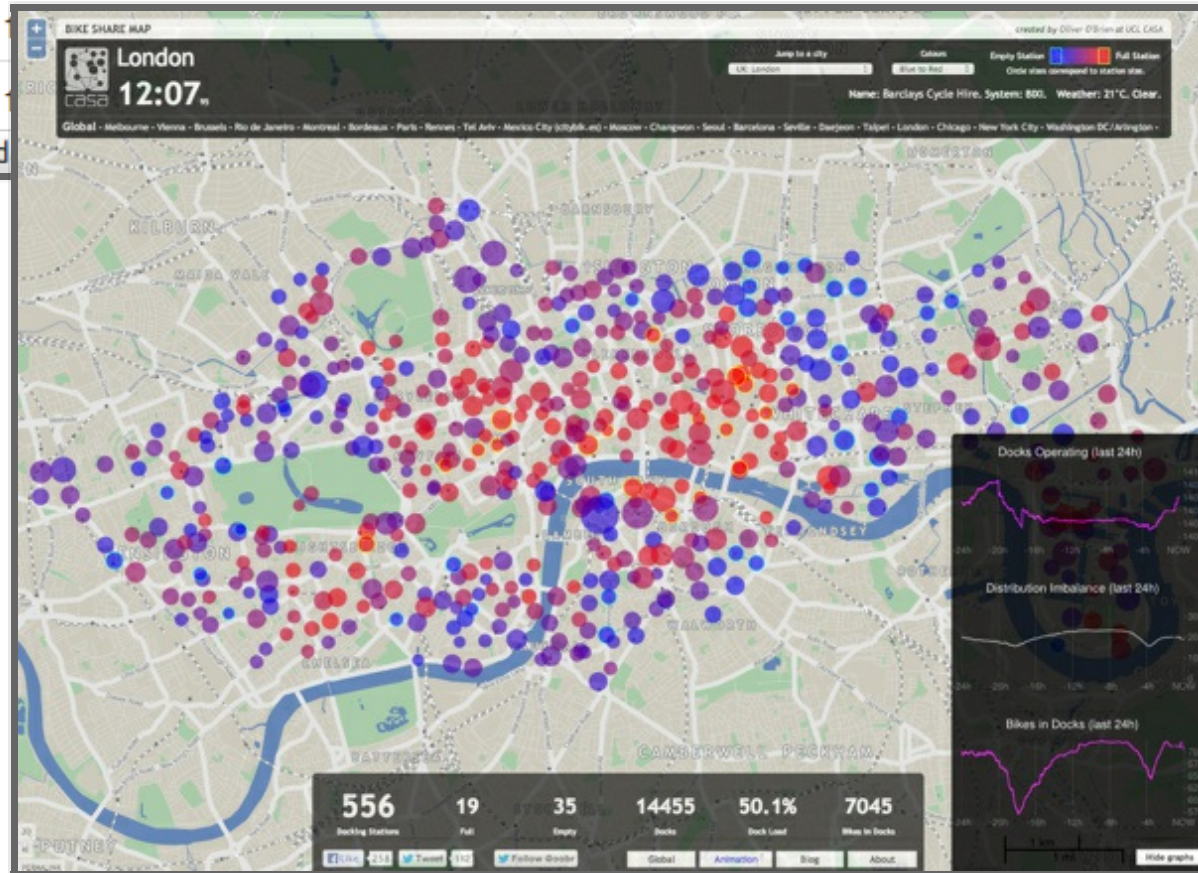
9 Taiwan

10 Australia

11 Chile

Example: BikeShare Map

global.census.okfn.org/d



Source

Nature

Contrast with traditional data

Class Quiz

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

- Geo-referenced tweets
- Land-registry house transaction values
- Google maps restaurant listing
- ONS Deprivation Indices
- Liverpool bikeshare service station status

Class Quiz - Answers

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

Challenges

Bias

- Traditionally, data used by urban researchers meets some quality standards (representativity, accuracy...)
- The accidental nature means new data sources will not always meet such standards
- This implies 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 even 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.

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, for instance)

Methods - Visualization

- Display of graphical summaries
- Arguably, not new to Geography, but more emphasis should be put on it
- Powerful to both obtain (explore the data) and communicate findings (tell stories with data)

Example: [Public Transit in Boston](#)

Methods - Machine learning

- Originated in computer science, blended with statistics
- Focus on prediction and pattern recognition
- Two main types of learning:
 - Supervised: present the computer some true relationships to "learn" a model, then use the model to infer others where no prediction is available (e.g. [Google flu trends](#))
 - Unsupervised: "let the data speak"... and the machine pick up the structure (e.g. [Livehoods](#))

New Vs Old? New + Old!

- Reconcile both worlds
- Complementary



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