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

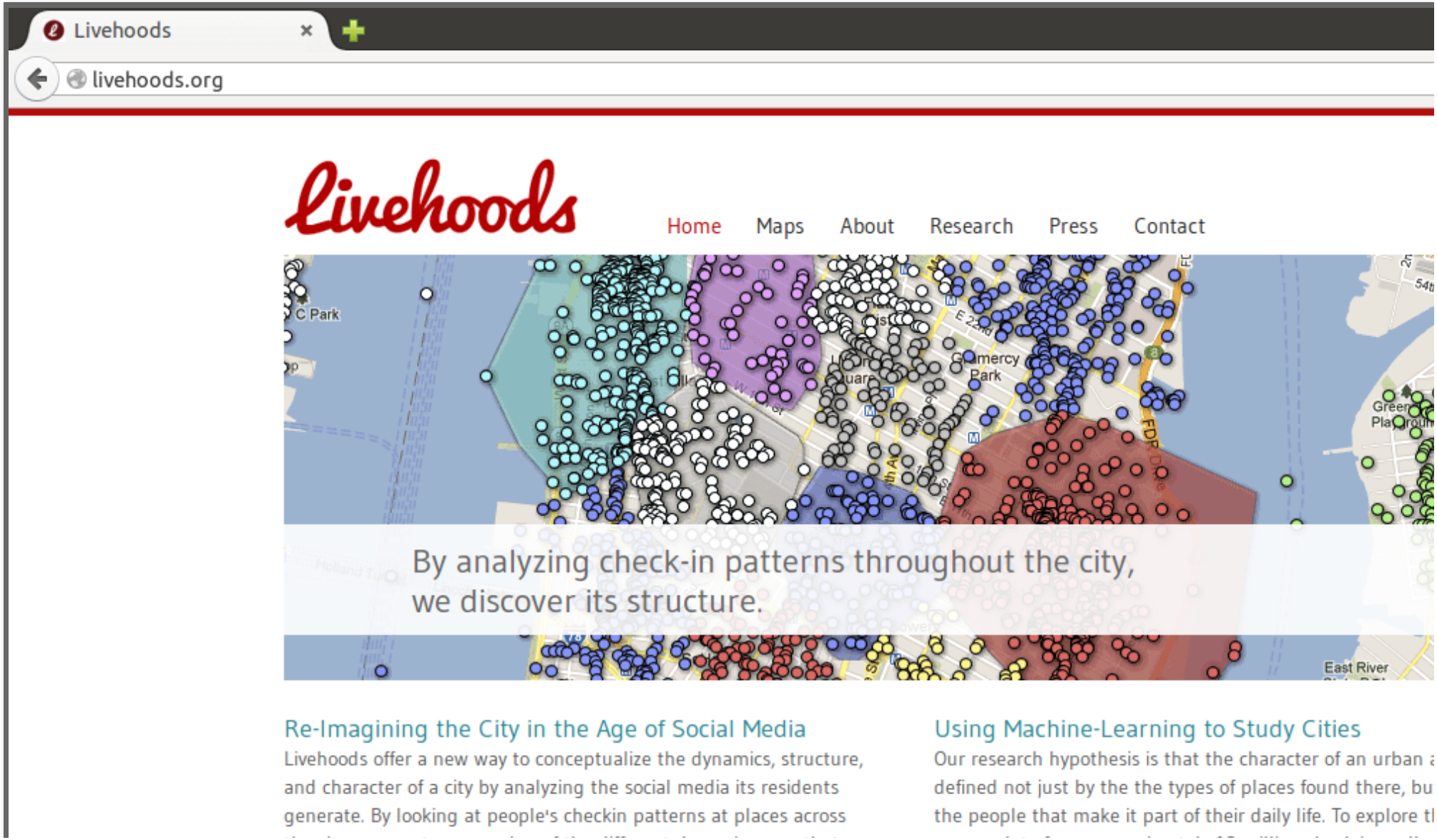
New sources of (spatial) data

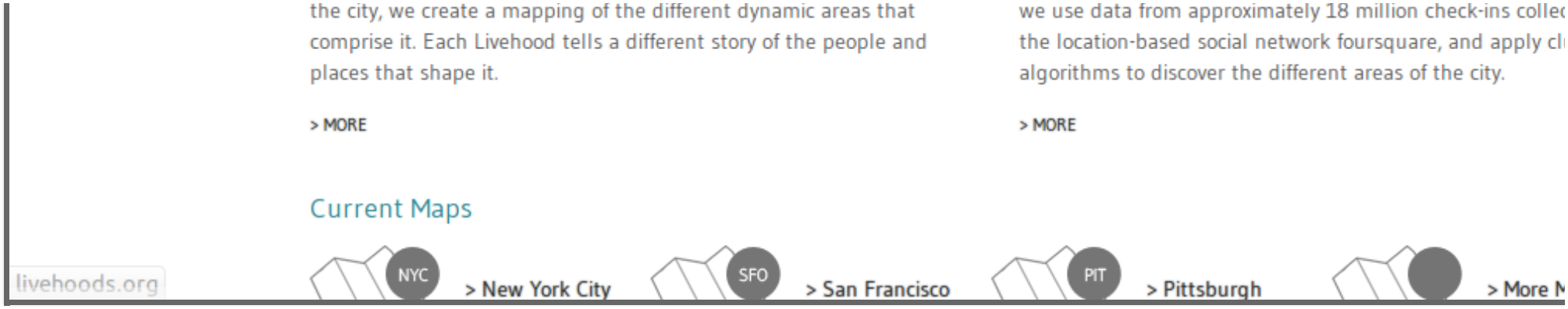
- Tied into the (geo-)data revolution
- Multi-type, but falling into same categories (polygons, points, surfaces...)
- Accidental --> very different nature
- Levels at which they are originated:
  - [Bottom up] "Citizens as sensors"
  - [Intermediate] Digital businesses/businesses going digital
  - [Top 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

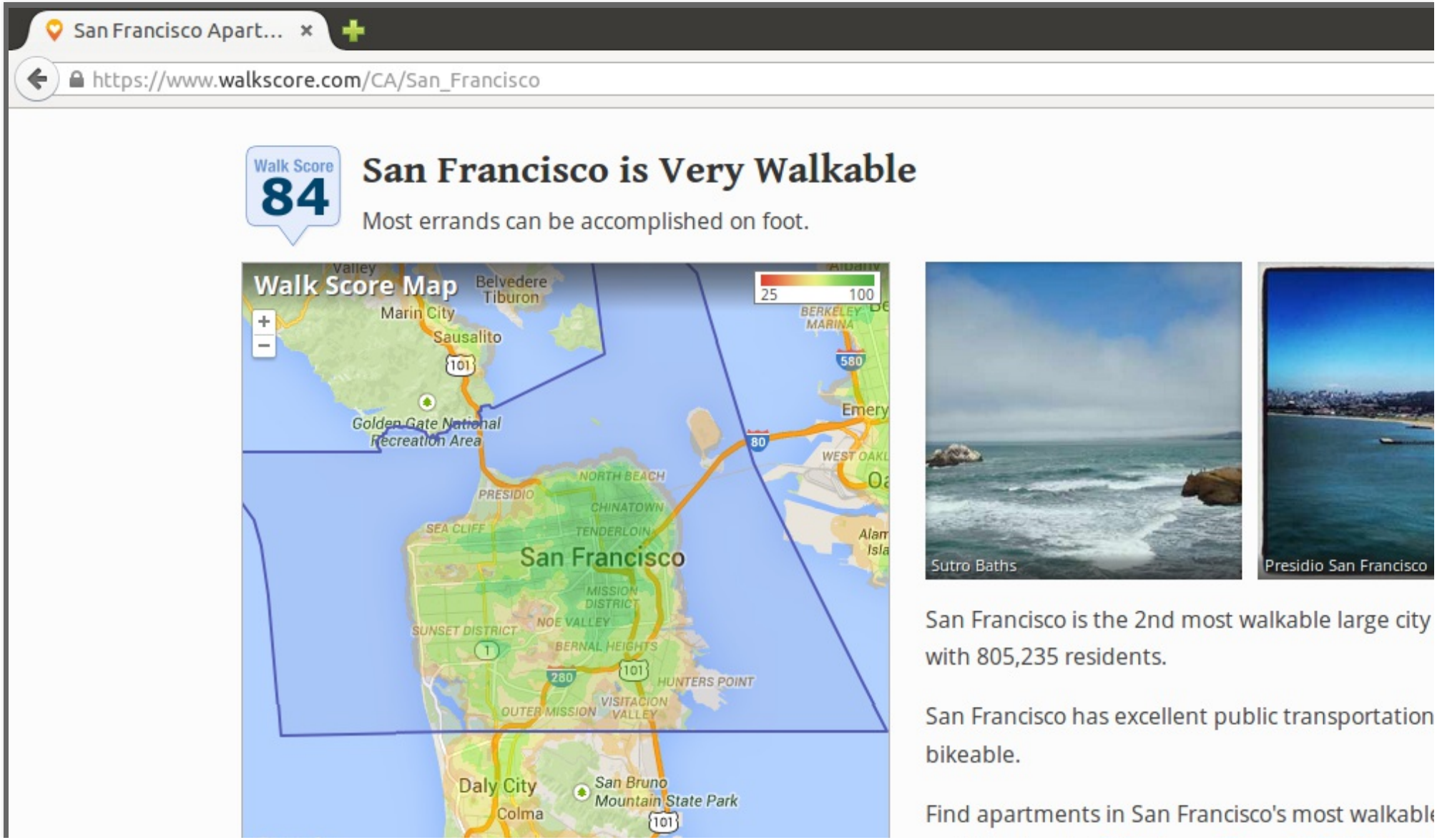


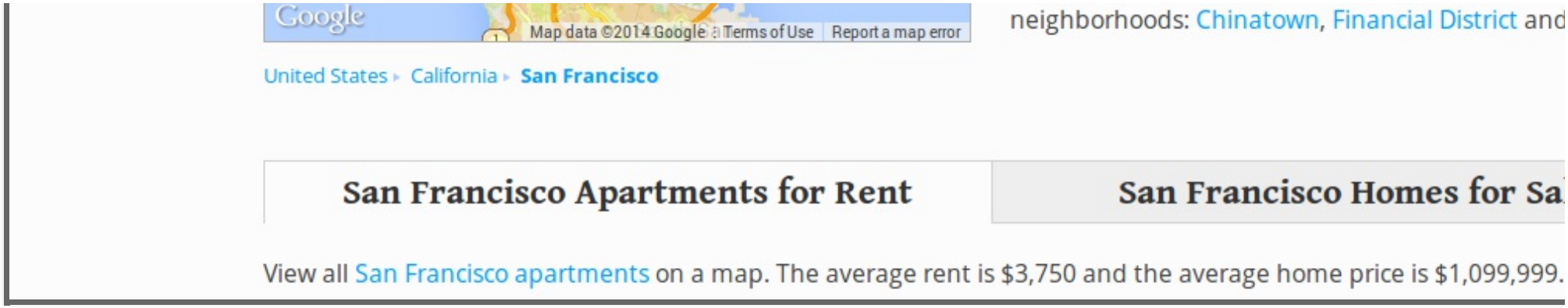


Business moving online

- Many of the elements and parts of bussiness 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, enirely 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



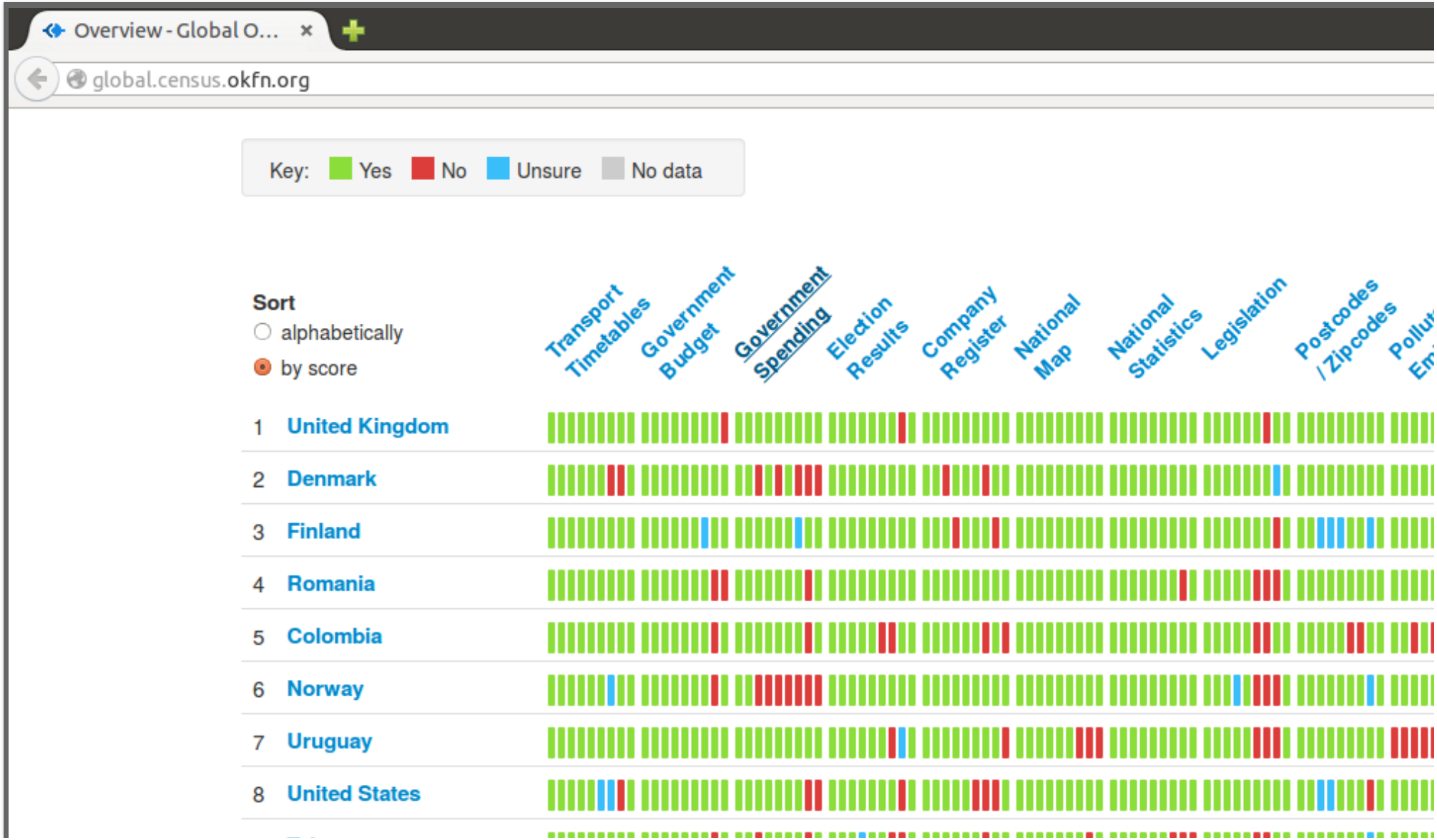


Open data for open governments

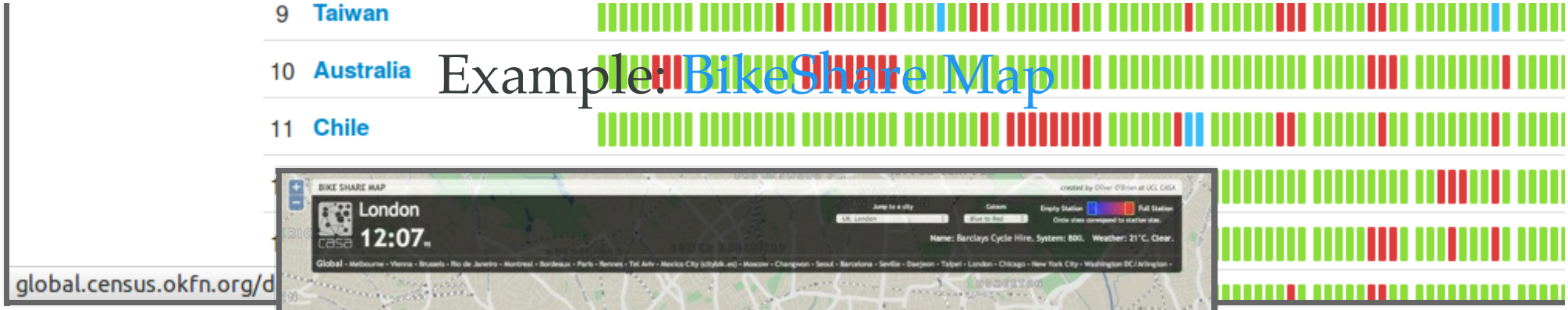
Government institutions release (part of) their internal data in open format. Motivations (Shadbolt, 2010):

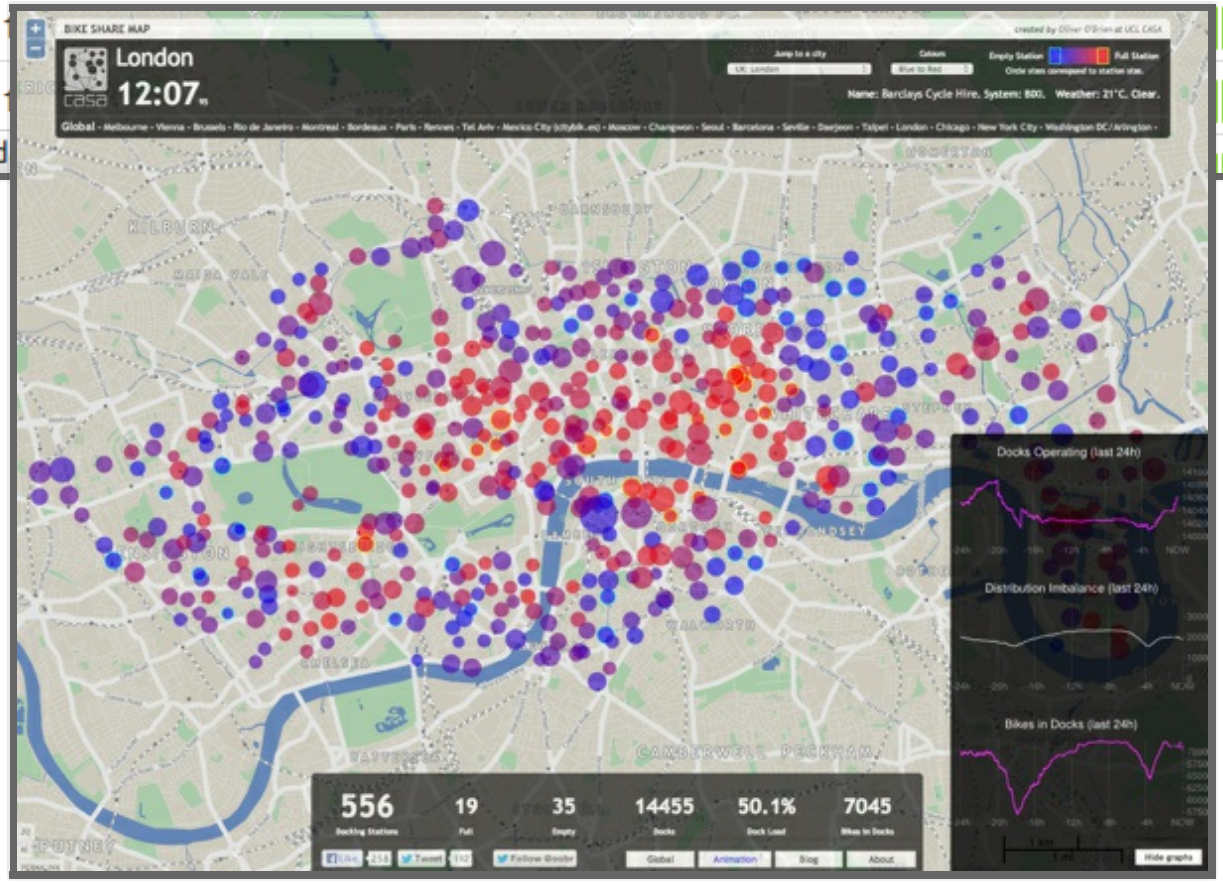
- Transparency and accountability
- Economic and social value
- Public service improvement
- Creation of new industries and jobs

Global Open Data Index 2014



Example: BikeShare Map





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