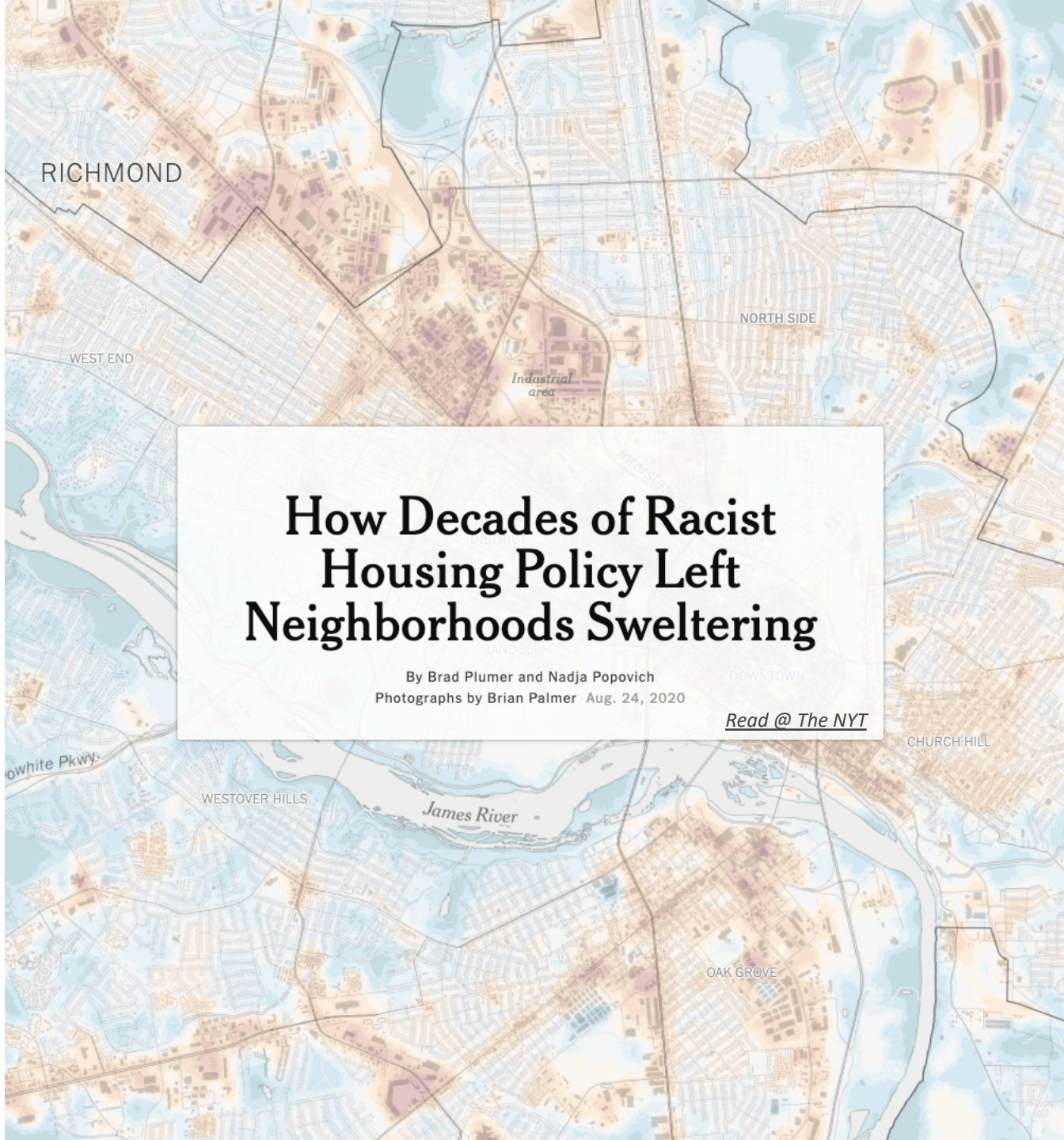
Introduction to *Urban* Data Science

Spatial and Urban Data

(EPA1316)

Lecture 2

Trivik Verma



Last Time

- Why Data Science?
- What is Data Science?
- Examine the **role** of evidence in policy
- Analyse **data** understanding and preparation **requirements**

Today

- What are data?
- Types of (geo-)data
- Traditional and new sources of spatial data
- Opportunities and Challenges
- New ways for traditional approaches

What are Data?

What are Data?

“A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data are multiple such measurements.”

Claim: everything is (can be) data!

Where do data come from?

- **Internal sources**: already collected by or is part of the overall data collection of your organization.
  For example: business-centric data that is available in the organization data base to record day to day operations; scientific or experimental data.
- **Existing External Sources**: available in ready to read format from an outside source for free or for a fee.
  For example: public government databases, stock market data, Yelp reviews, [your favorite sport]-reference.
- **External Sources Requiring Collection Efforts**: available from external source but acquisition requires special processing.
  For example: data appearing only in print form, or data on websites.

Ways to gather Online data

How to get data generated, published or hosted online:

- **API (Application Programming Interface):** using a prebuilt set of functions developed by a company to access their services. Often pay to use. For example: Google Map API, Facebook API, Twitter API
- **RSS (Rich Site Summary)**: summarizes frequently updated online content in standard format. Free to read if the site has one. For example: news-related sites, blogs
- **Web scraping**: using software, scripts or by-hand extracting data from what is displayed on a page or what is contained in the HTML file (often in tables).

Web Scraping

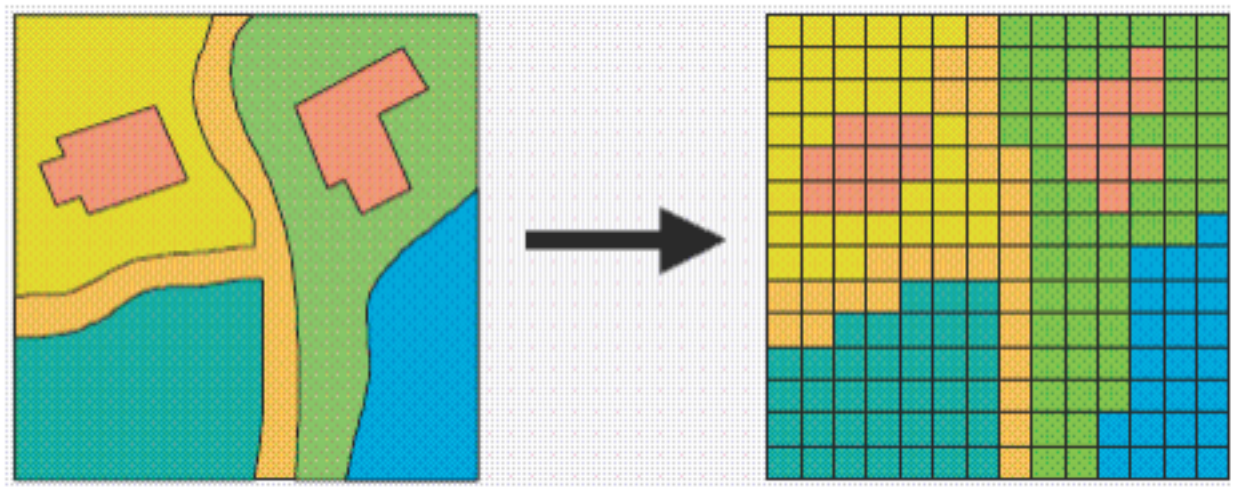
- **Why do it?** Older government or smaller news sites might not have APIs for accessing data or publish RSS feeds or have databases for download. Or, you don't want to pay to use the API or the database.
- **How do you do it?** (beautifulsoup / python package) – advanced material in lab 2
- **Should you do it?**
  - You just want to explore: Are you violating their terms of service? Privacy concerns for website and their clients?
  - You want to publish your analysis or product: Do they have an API or fee that you are bypassing? Are they willing to share this data? Are you violating their terms of service? Are there privacy concerns?

Representing the world digitally

GIS Data

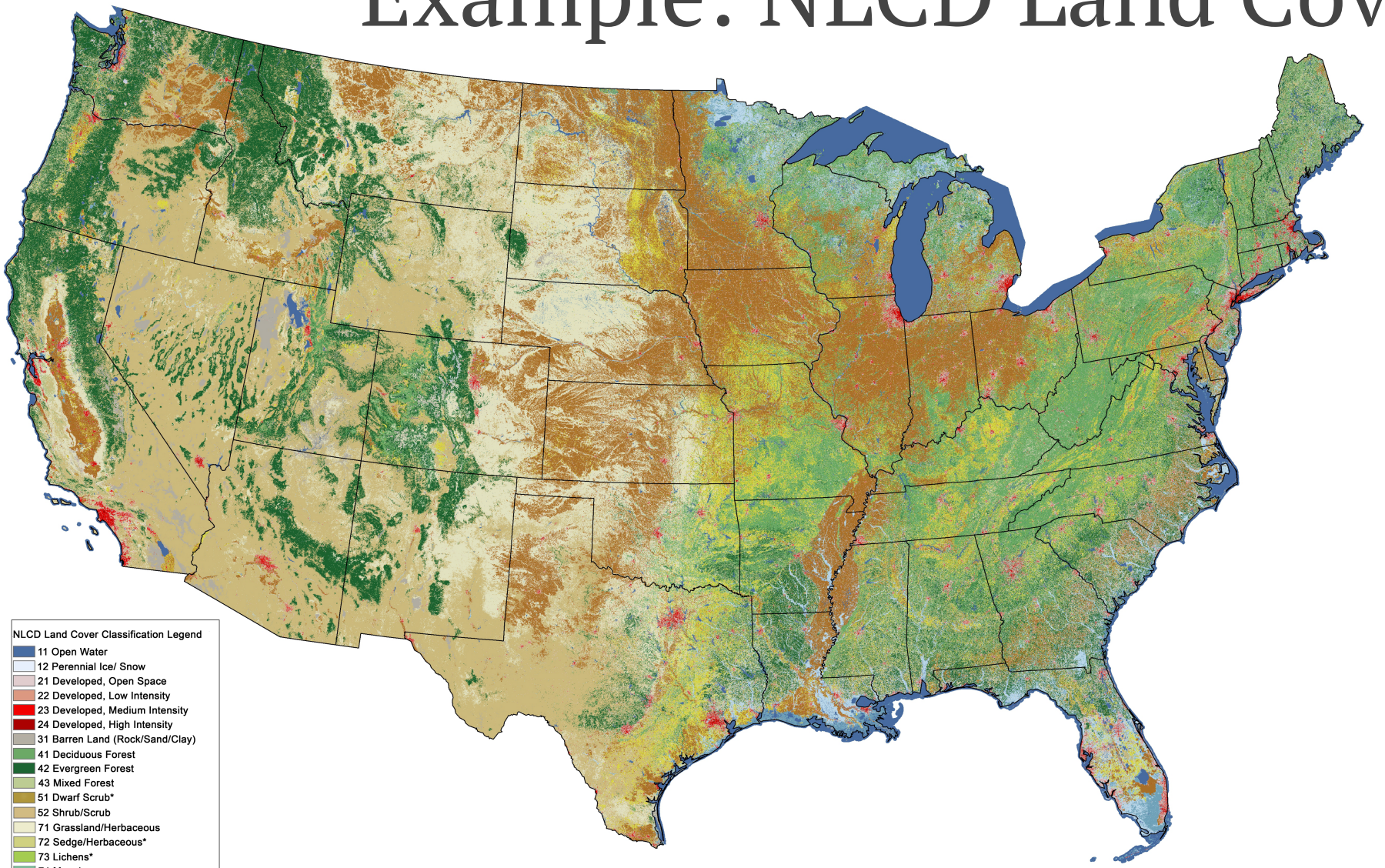
Traditionally, geographic information is represented as:

- Vector finite set of entities (shapes/geometries)
- Raster images encoding surfaces (values, colours, etc.)



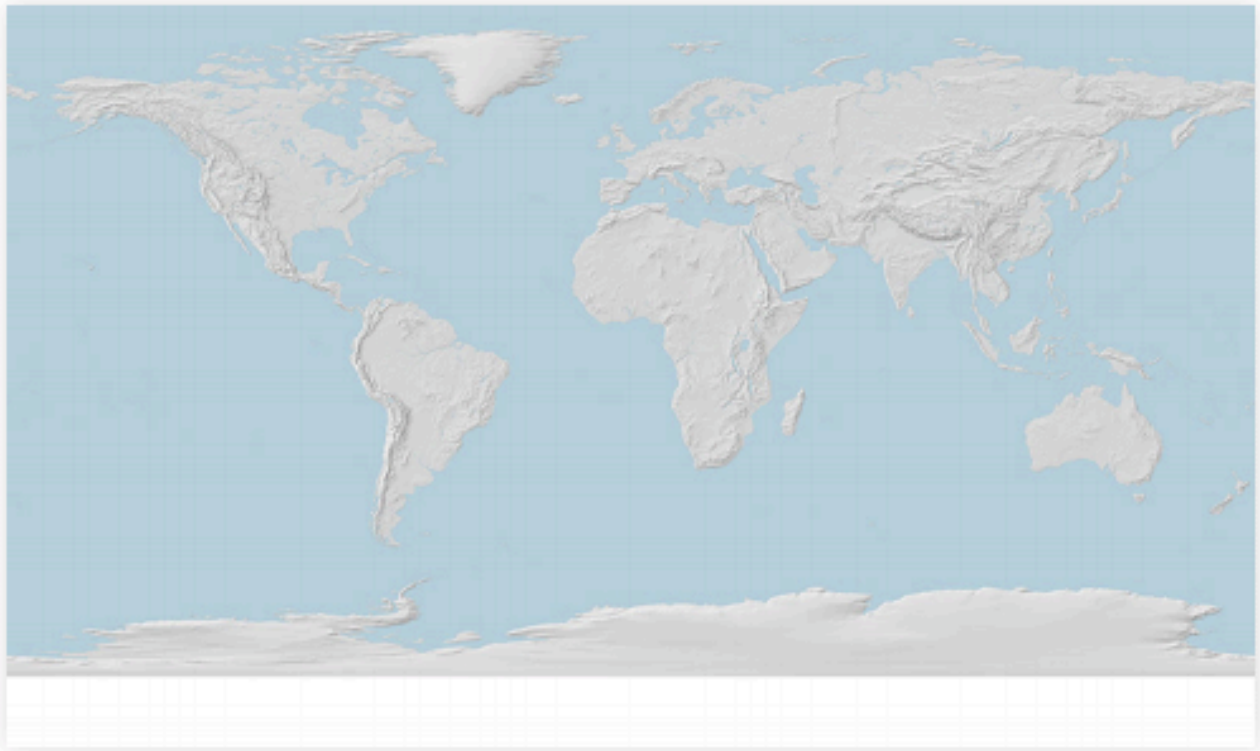
One of the most common types of raster data is land cover derived from satellite imagery. Land-cover data is produced by assigning each pixel in a Landsat thematic mapper image to one of 16 land-cover classes using a procedure known as unsupervised classification.

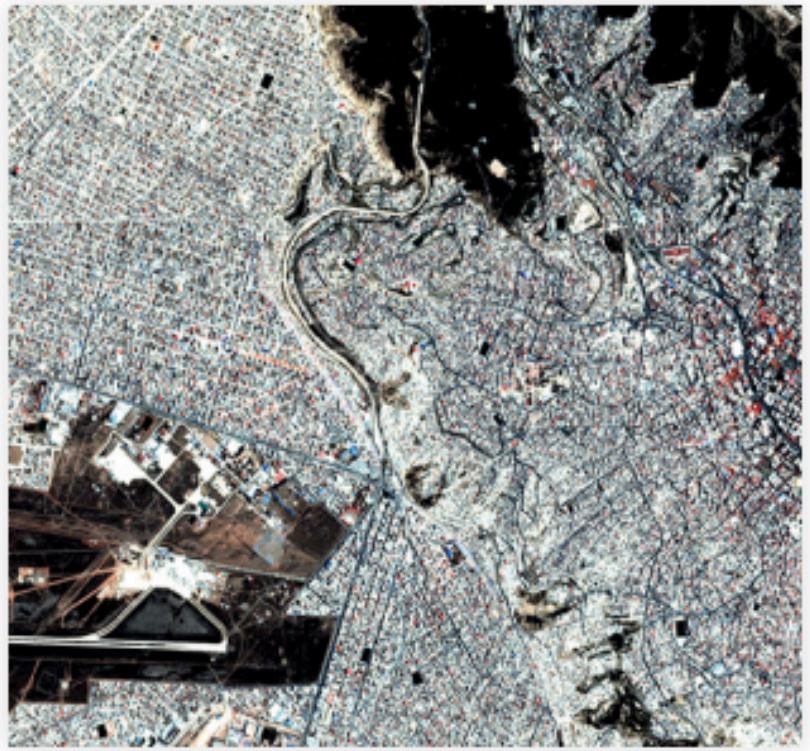
Example: NLCD Land Cover



Vector

Raster





Good old Spatial Data

[source]

Good old 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 Data (-)

But also:

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

Examples

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

Break

WATER

WALK

COFFEE OR TEA

MAKE FRIENDS

New sources of *Spatial* Data

New Sources of *Spatial* Data

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 and 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 (The Hague Cijfers)

Class Quiz

Class Quiz

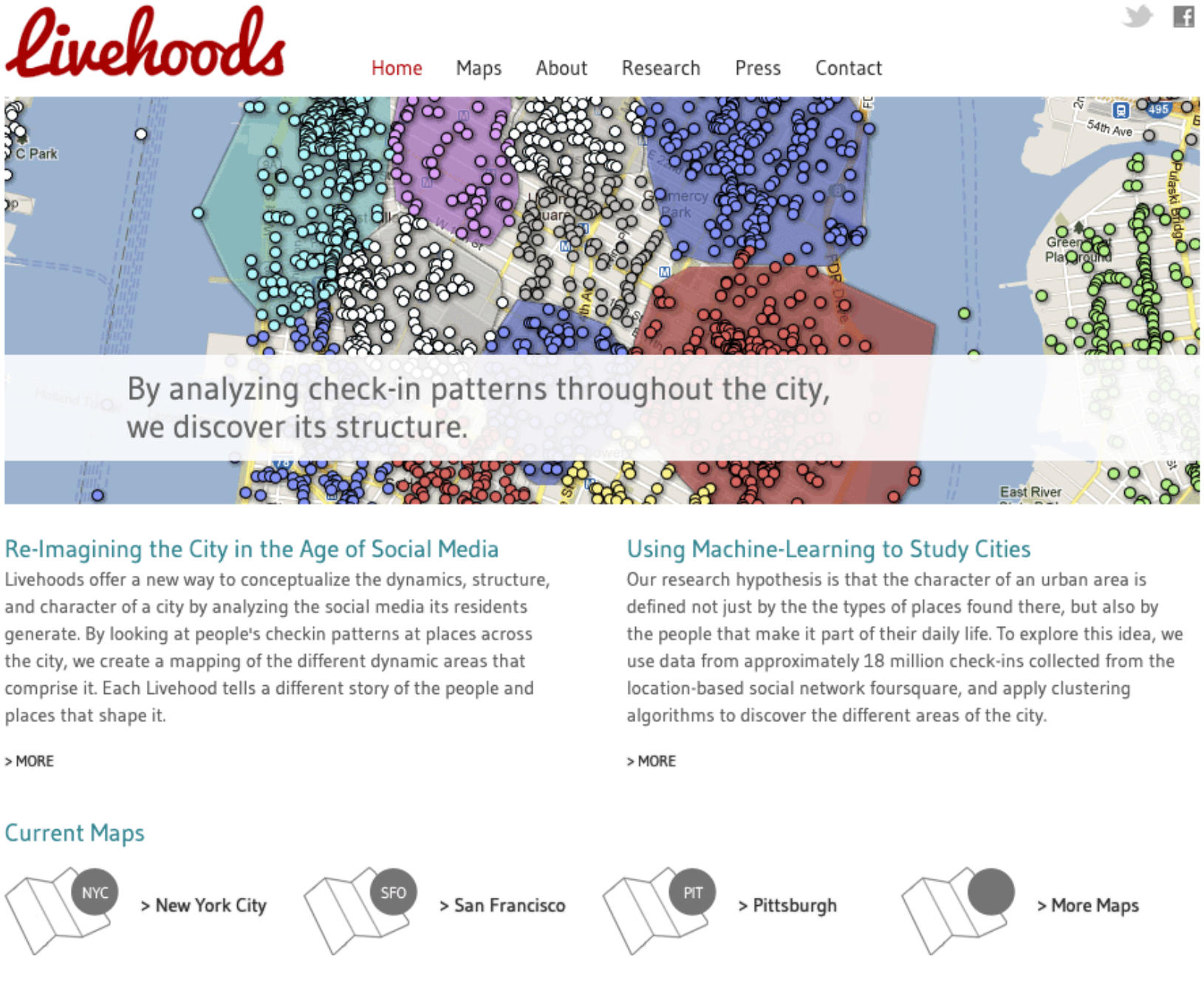
What is the origin of the following sources of (geo-)data:

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

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



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

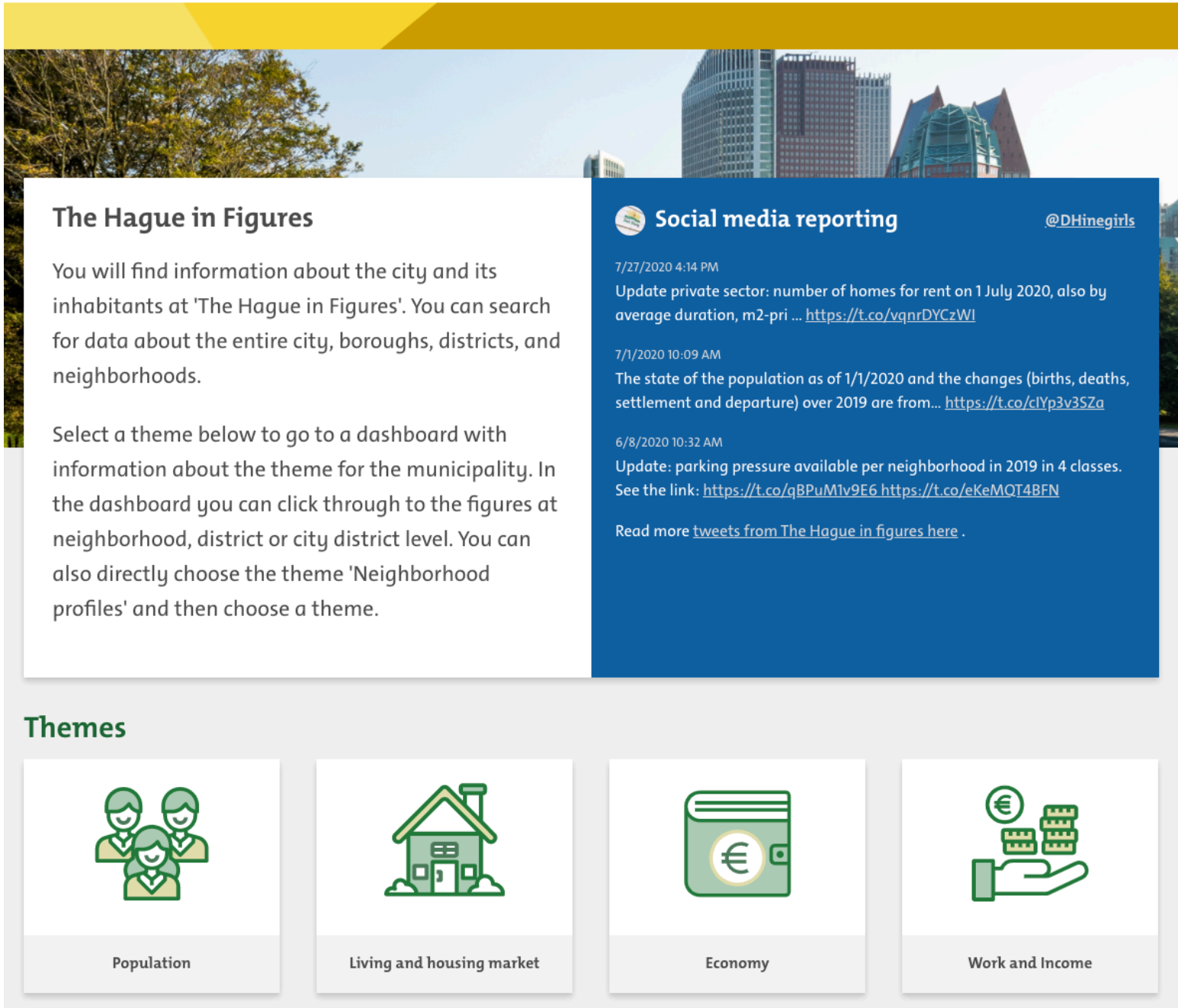
Example:
Walk Score

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

Example:
The Hague Cijfers



Opportunities and Challenges

Opportunities (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 (Arribas-Bel, 2014)

- Bias
- Technical Barriers
- Methodological “mismatch”

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 *difficult* to access
- 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 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 or any other discipline –> borrow techniques from other disciplines, or even create new ones

New + Old

Traditional data:

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

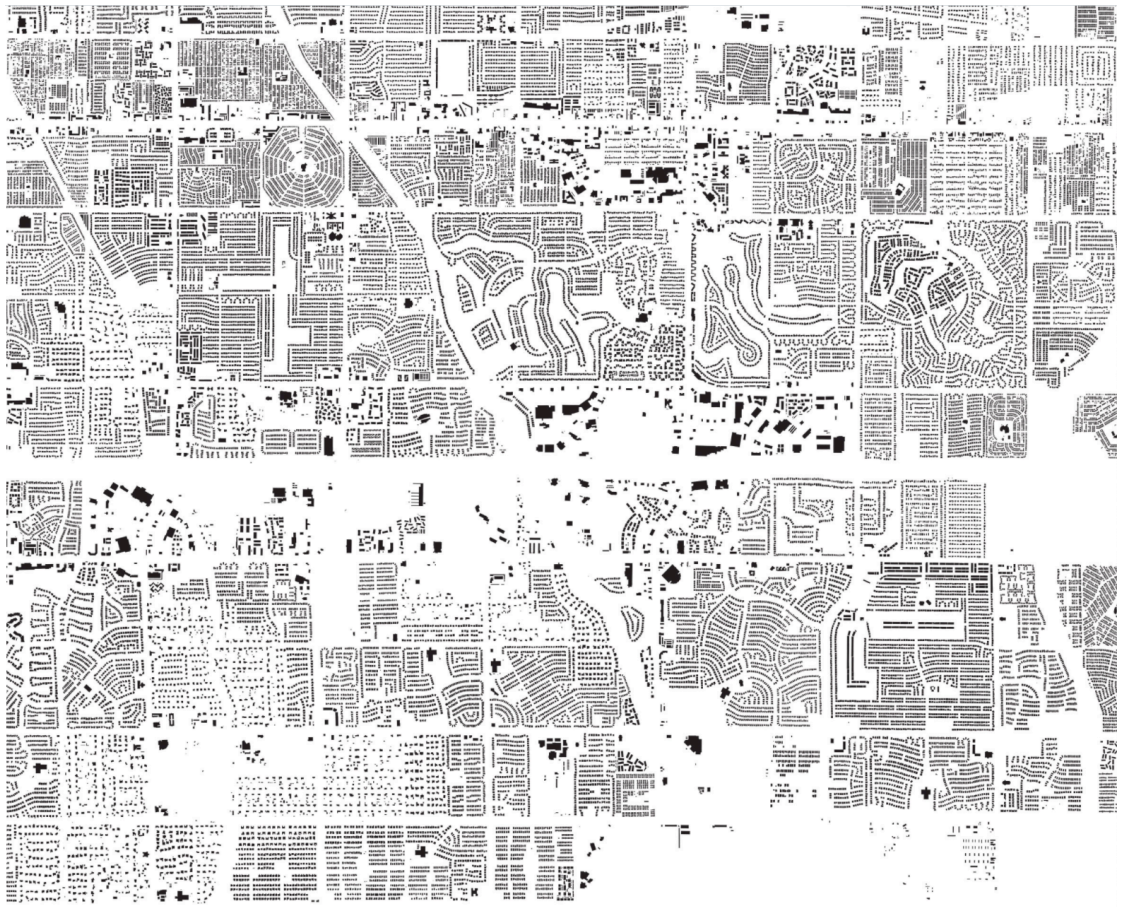
Accidental data:

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

Old/New, raster/vector . . .

Traditional approaches to represent the world in a computer are blending thanks to new forms of data

Keep an open mind to tools, approaches, and methods



A NATION OF SUBURBS

Mesa, Ariz. America's suburban streets twist and flow, with their wild involutions and curving cul-de-sacs. Mesa's suburbs are especially imaginative, particularly from above. The feeling of meandering through a place whose layout is designed to thwart speed and comprehension is familiar to anyone who, in the days before GPS, needed to pick up a friend or deliver a pizza in an unfamiliar neighborhood.

[source]

EUROPEAN COMMISSION

Global Human Settlement

European Commission > EU Science Hub > GHSL

Home | About | Copernicus | GEO | Documents | Atlases | Applications | Degree of Urbanisation | Data | Tools | Visualisation | News

GHSL - Global Human Settlement Layer

A new open and free tool for assessing the human presence on the planet

- Produces new global spatial information, evidence-based analytics and knowledge describing the human presence on the planet
- Operates in an open and free data and methods access policy (open input, open method, open output)
- Supported by the Joint Research Centre (JRC) and the DG for Regional and Urban Policy (DG REGIO) of the European Commission, together with the international partnership GEO Human Planet Initiative GEO GROUP ON EARTH OBSERVATIONS

News ***26/03/2020*** ***Call for Contribution to the JRC Atlas of the Human Planet 2020*** *it will showcase applications of the GHSL data. Go to our* ***news page*** *for the details*

https://ghsl.jrc.ec.europa.eu/index.php

For next class..

Finish Lab 02 to practice programming

Submit Homework 02 for peer review on Brightspace

Check Assignment 1 – due in **Week 3** on Friday at **2330**

See "To do before class" for every lecture (~ 1 hour of self study)

Read paper for **Discussion** session before every Friday

Post questions on the **Discussion** forum on Brightspace (especially about **Data**)