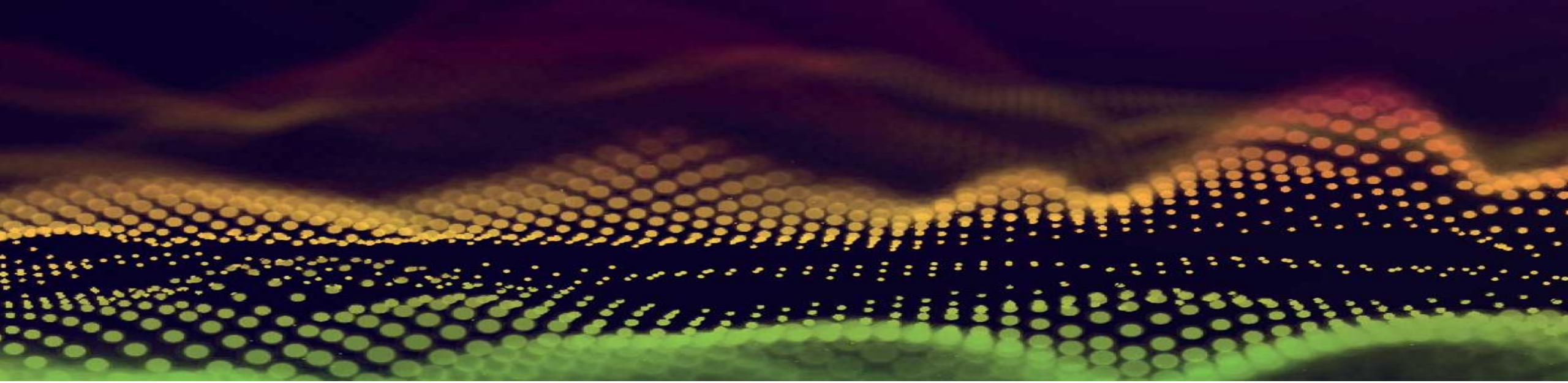


Social Data Science needs Data

Andrew McHugh, Urban Big Data Centre, University of Glasgow

For CREATe's ***Information, (research) data and open science*** Workshop

October 10th, 2019

Urban Big Data Centre

JOINTLY FUNDED BY E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

University of Glasgow

About the Urban Big Data Centre

- UK Government (Economic and Social Research Council) funded
- Priorities:
  - Data infrastructure and collections
  - Priority research strands: transport & mobility; neighbourhood, housing & environment; education, skills & productivity; big data & urban governance
  - Combining social science research with data analytics and computing science
- Overall aims:
  - Achieve public policy impact
  - Critically evaluate role and value of big data and urban analytics
  - Enhance data and methods

"Promoting innovative research methods and the use of big data to improve social, economic and environmental well-being in cities"

Urban Big Data Centre

Urban data context – data sources

Urban Big Data	Examples
Sensor systems	Environmental, water, transportation, building management sensor systems; connected systems; Internet of Things
User-Generated Content	Participatory sensing systems, citizen science projects, social media, web use, GPS, online social networks and other socially-generated data
Administrative (governmental) Data	Open administrative data on transactions, taxes and revenue, payments and registrations; confidential person-level microdata
Private Sector Data	Customer transactions data from store cards and business records; fleet management systems; usage data from utilities and financial institutions; product purchases and terms of service agreements
Arts and Humanities Data	Repositories of text, images, sound recordings, linguistic data, film, art and material culture, and digital objects, and other media
Hybrid Data Sources and Synthetic Data	Linked data including survey-sensor, census-administrative records

Thakuriah, P., N. Tilahun and M. Zellner (2017). Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery. In *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics*, Springer, NY, pp. 11-48.

Urban Big Data Centre

Contexts for our Work

- **Technological**
  - Information generation & capture
  - Data management
  - Data processing
  - Dissemination and discovery
- **Methodological**
  - Data preparation (IR, extraction; linkage; cleaning, anonymisation, quality)
  - Data analysis (methods to analyse domain challenges, uncertainty and bias)
- **Theoretical and Epistemological**
  - Domain knowledge – metrics, definitions, ideologies
  - Understanding limits of data driven approach
  - Information paradoxes
- **Political Economy**
  - Data entrepreneurship, innovation networks and power structures
  - Value propositions
  - Data acquisition and availability
  - Privacy, security and information governance
  - Responsible innovation

Typical Research / Approaches

- Urban metabolism – **real time analytics** using **social media** and **GPS data** to identify spatio-temporal activity clusters (functional usage / stay duration…) and semantically annotated to connect land use PoI and transport networks
- **Geolocalisation** of social media data – identify under-reported phenomena such as road traffic incidents, and explore relationship between crashes and crime
- **Wearable sensors** combined to show mobility patterns and behaviours (indoor walking; social exclusion; travel modality)
- **Computer vision methods** to extract pedestrian / other object counts from CCTV sensor networks
- Validating and informing infrastructural investments using resources like **Strava Metro activity data**

Key challenges

Skills capacity

Resources

Data sharing

Trust

Urban Big Data Centre

Data acquisition – a major challenge

- Getting hold of data to support our research takes up a lot of my time

- There are many reasons why organisation **do not** make their data available
  - Lack of capacity to do so
  - Perceptions of risk (e.g. incompatibilities with privacy legislation)
  - Perceived conflict with existing business models
  - Fears that research results will reflect badly on them, expose them to criticism or be received negatively by other stakeholders

- Conversely, there are several reasons why it **might be a good idea**
  - Establish credibility of data as robust evidence base
  - Facilitate potentially beneficial analysis
  - Often datasets are by definition publicly owned
  - Transparency just might be a good thing

Urban Big Data Centre

Case study – AirBnB (1)

- AirBnB is an online marketplace for arranging or offering lodging, homestays or tourism experiences
- What are the impacts of the rapidly growing sharing economy for private rented sector property market?
  - AirBnB is not enthusiastic about facilitating such research
  - As an unregulated sector there is little reliable data available
  - Some third party providers make these data available (under US copyright law) but problems for robust academic work:
    - Sampling limitations
    - Data quality issues
    - Black box models

Case study – AirBnB (2)

- We want to automate the **scraping** of AirBnB on a systematic basis to capture and store publicly accessible web content and facilitate research
- Aiming to leverage copyright text and data mining exception to do so legally
- Data collection designed to enable us to demonstrate empirically and reliably:
  - Scale and growth of AirBnB
  - Spatial change and focus of short-term lets
  - Occupancy
  - Price
  - Availability

Urban Big Data Centre

Case study – AirBnB (3)

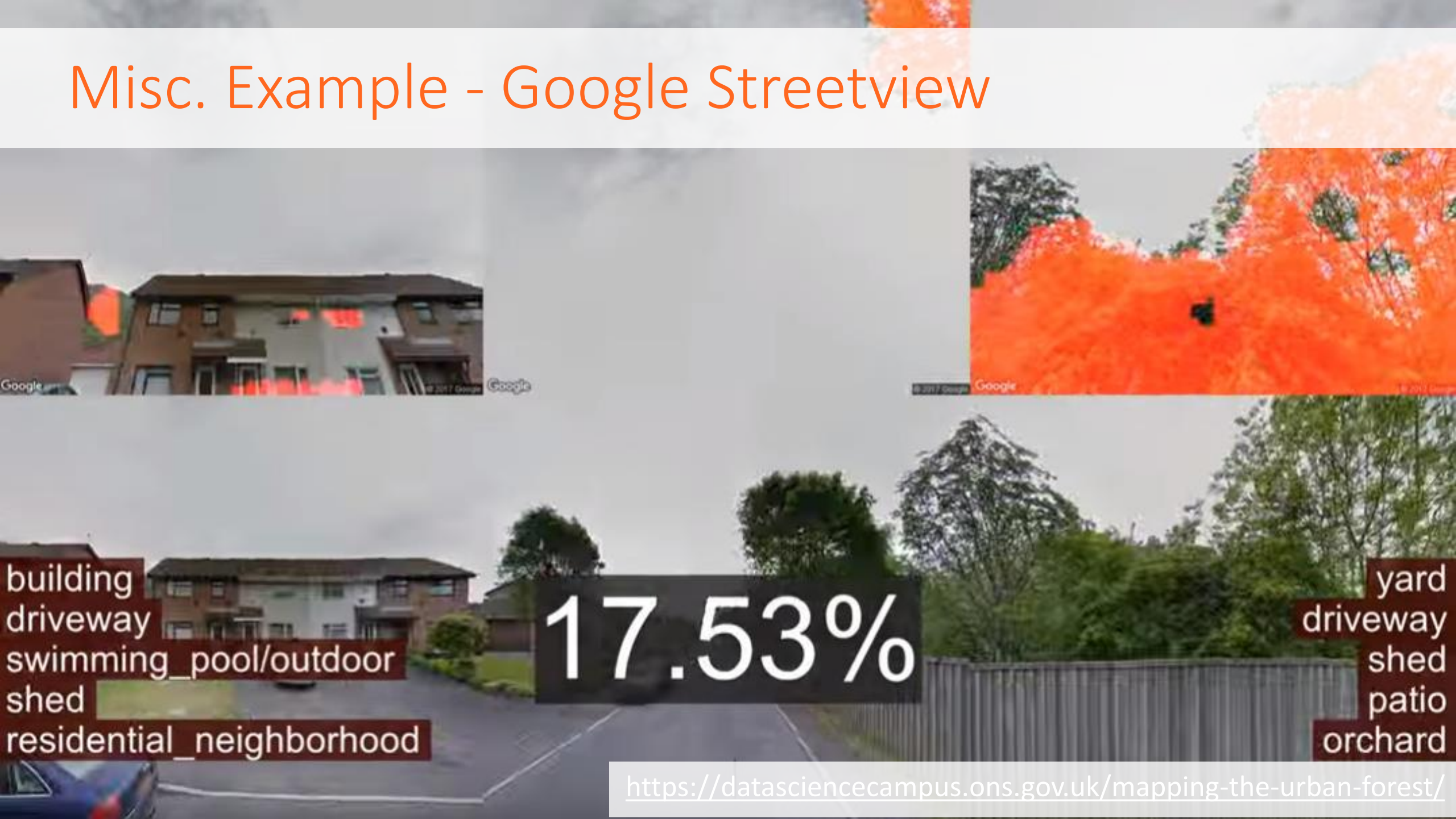
- We plan to collaborate with colleagues in Computing Science and Business / Management areas to explore:
  - What impact does AirBnB have on the availability of private renting stock?
  - Where is that impact greatest?
  - What is the relationship with areas of social deprivation? Does the rise of the sharing economy increase inner-city gentrification and the suburbanisation of poverty?
  - Are AirBnB properties falling below the standards for the PRS e..g in terms of occupancy levels?
  - What impact does AirBnB have on the existing hospitality industry?

Urban Big Data Centre

Case study – AirBnB (4)

- Scraping means automating the transfer of data from a web site
  - Text pattern matching
  - HTTP programming
  - HTML / DOM parsing
  - Computer vision
- Unsupported by the web content owner, highly prone to break if the website changes
- Tools are available to facilitate, including specifically for AirBnB
- Negotiate website controls intended to block nuisance hosts – employing anonymous web proxy hosts – but what do we know about these proxies?

Misc. Example - Google Streetview



https://datasciencecampus.ons.gov.uk/mapping-the-urban-forest/

Misc. Example - Scottish EPCs

Ministry of Housing, Communities & Local Government

Energy Performance of Buildings Data: **England and Wales**

Access to Energy Performance Certificates and Display Energy Certificates data for buildings in England and Wales. Searchable, browsable and downloadable individually or in bulk. Register now for access to the data.

For householders
Check the data for your property

For researchers
An indication of energy use

For business
Create new products and services

For policymakers
Make data driven decisions

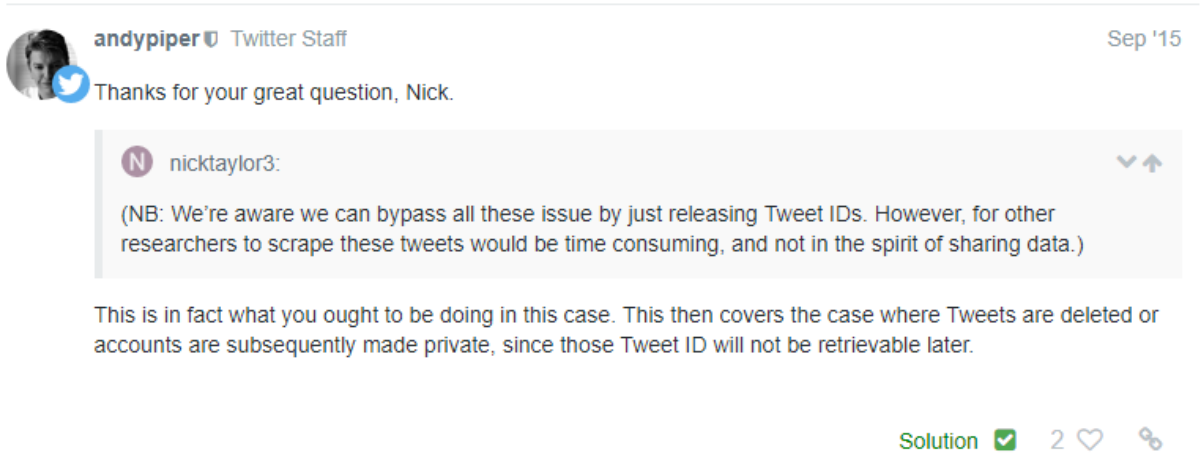
Contrasting situation between on one hand, England and Wales (open data) and Scotland (restricted by Ts and Cs, Captchas and Watermarks)

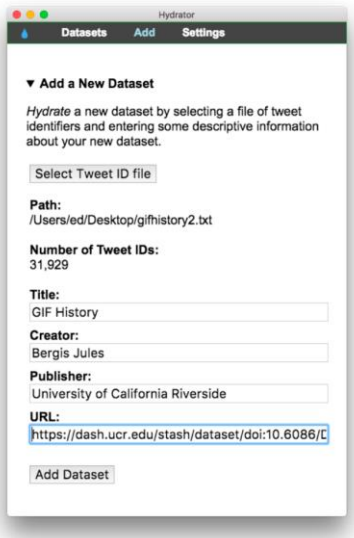
- You must not use or attempt to use any automated program to access the Scottish Energy Performance Certificate Register or this web site, or to search, replicate display or obtain links to any part of this site. All access to and use of the services via any automated software agent is prohibited. This includes without limitation, any mechanical program, screen scraper, spider or other web crawler.

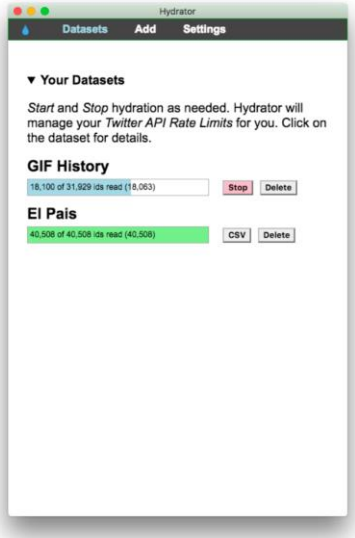
Urban Big Data Centre

Misc. Example - Twitter data collection

- Twitter has several means for automating the capture of online data in bulk (e.g. Streaming API, albeit limited to 1% of tweets)
- It also provides explicit permission to share for academic research purposes – but limits to IDs only:







- https://twittercommunity.com/t/twitter-and-open-data-in-academia/51934/4
- Questions surround privacy – e.g. "off-Twitter matching" and managing other peoples' sensitive information

Urban Big Data Centre

Misc. Example - Adzuna

- Interest in changing skills and employment landscape
- Terms permit academic research use and RESTful API facilitates this
- Whose responsibility is it to make sure 3rd party content owners rights are respected?

Urban Big Data Centre

Summary / Thoughts

- The technical landscape is complex – but we know that quite well
- Real scope for innovation around legal and regulatory issues – perceived copyright, licensing, privacy risk can have quite **a paralysing effect** for responsible researchers
- But then other issues are really relevant too:
  - Ethics
  - Relationship management
  - Practical aspects of managing

Urban Big Data Centre

Thanks – Questions?

To stay up to date with our research and other activities you can sign up to our newsletter on our website and follow us on social media:

www.ubdc.ac.uk

/urbanbigdata

@UrbanBigData

Urban Big Data Centre

Urban Big Data Centre

JOINTLY FUNDED BY

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

University of Glasgow