Challenges in Geocoding Socially-Generated Data

Jonny Huck (2nd year part-time PhD student)

Duncan Whyatt

Paul Coulton

LANCASTER UNIVERSITY

Lancaster Environment Centre

School of Computing and Communications

Royal Wedding

- Whilst we were all on the way back from GISRUK at Portsmouth last year, Prince William and Kate Middleton got married.
- c.**1.7 million** Tweets collected worldwide.
- Emotive
- Predictable Timescale

Socially-Generated Data

- Data created within social networking websites (Twitter, Facebook etc.).
- Potentially a rich dataset.
- Significant growth in use as geographical data.
- c.1% has coordinates already attached.
- Most data will need to be geocoded, using the **place name** specified in the profile of the user.

Geocoding

- Adding spatial information, to non-spatial data.
- Both **coordinates**, and **address components**.
- Formerly the domain of skilled specialist operators.
- This changed with free, online **global** geocoding services.
- **Multiple results** often returned.

Aims and Objectives

- Highlight the issues that we have found.
- Explore the **impact** that this can have upon analysis.
- Suggest a **methodology** to attempt to address both of these issues.
- Investigate the effect that applying these techniques can have upon analysis.

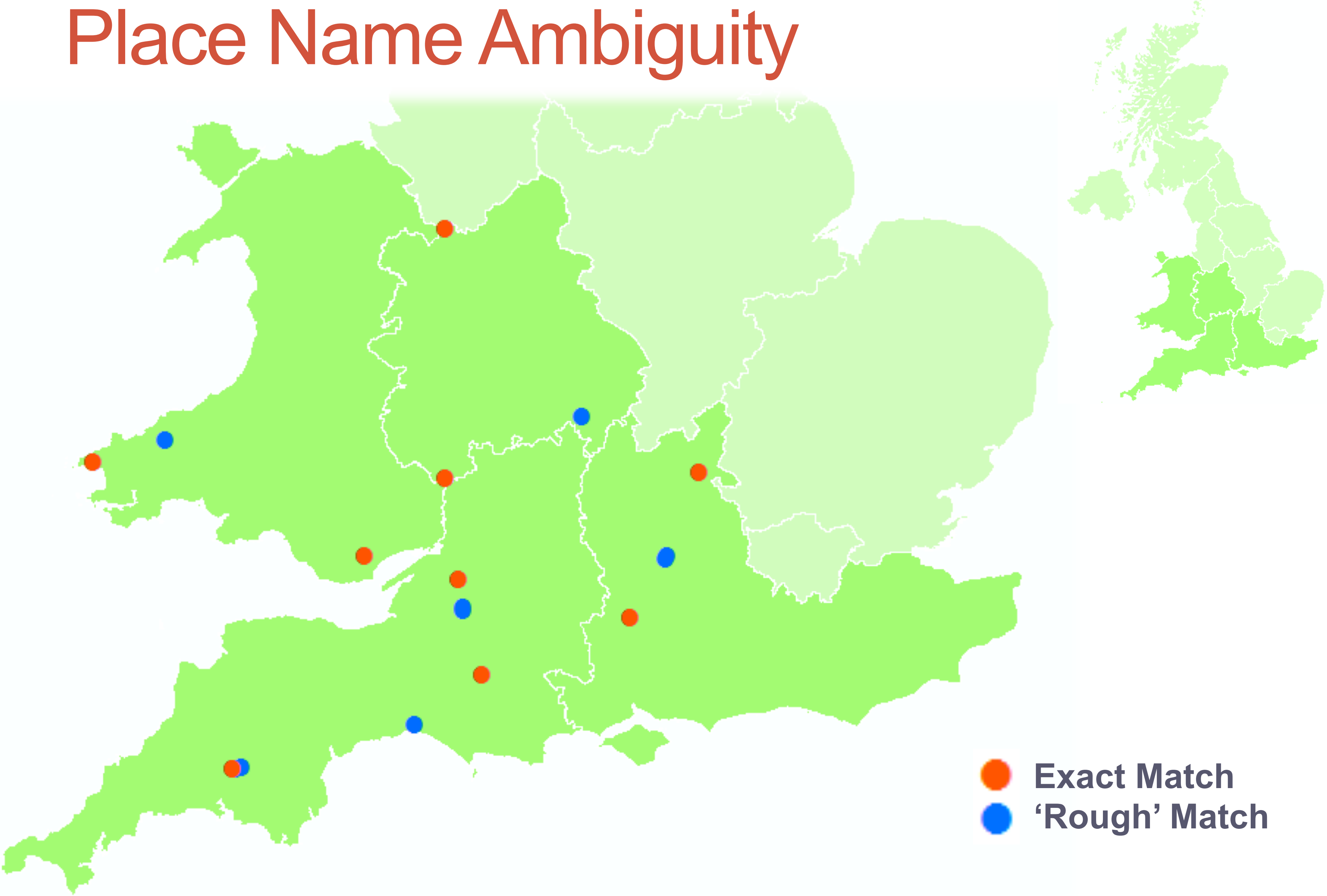
Problem 1:

Place Name Ambiguity

Place-Name Ambiguity

- Place-names are **not unique** identifiers.
    - Multiple places have the **same name**.
    - A single place can have **multiple names**.
- Automated identification of the ‘correct’ place is therefore un-reliable.

Place Name Ambiguity

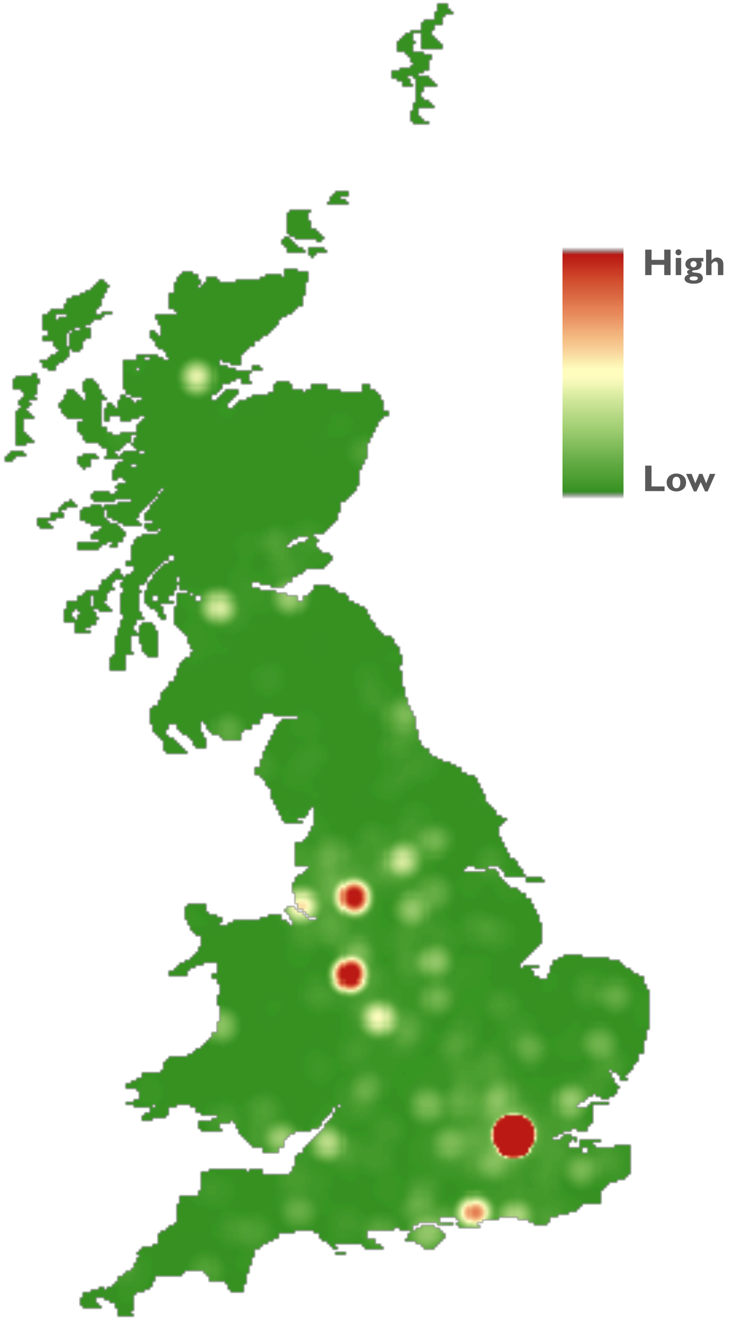


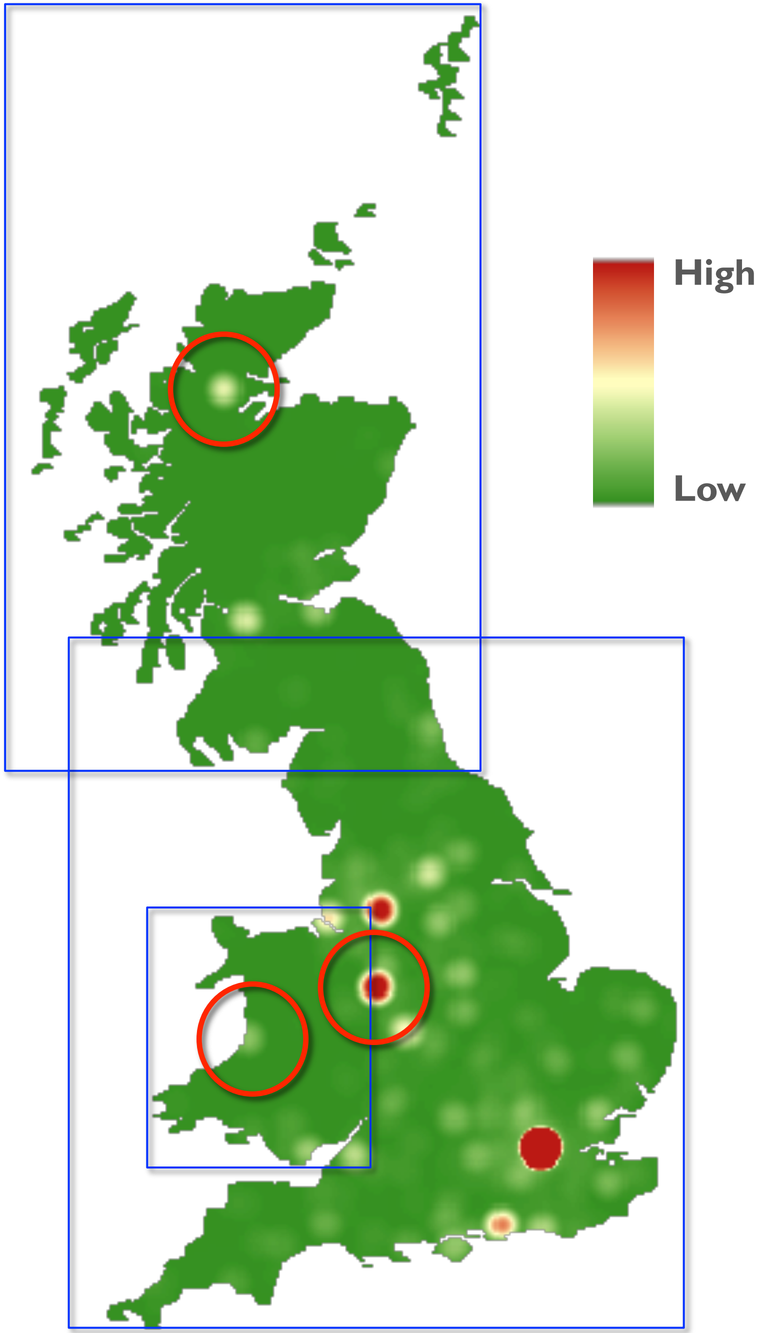
Problem 2:

Undefined Level of Detail

Undefined Level of Detail

- Comparison of data at a multitude of **levels of detail** within the same analysis.
- ‘**False Hotspots**’ occur at the centroid of places.
  - Creates the false impression of activity
  - Masks variations in actual activity.


High
Low


High
Low

Methodology 1:

Place Name Ambiguity

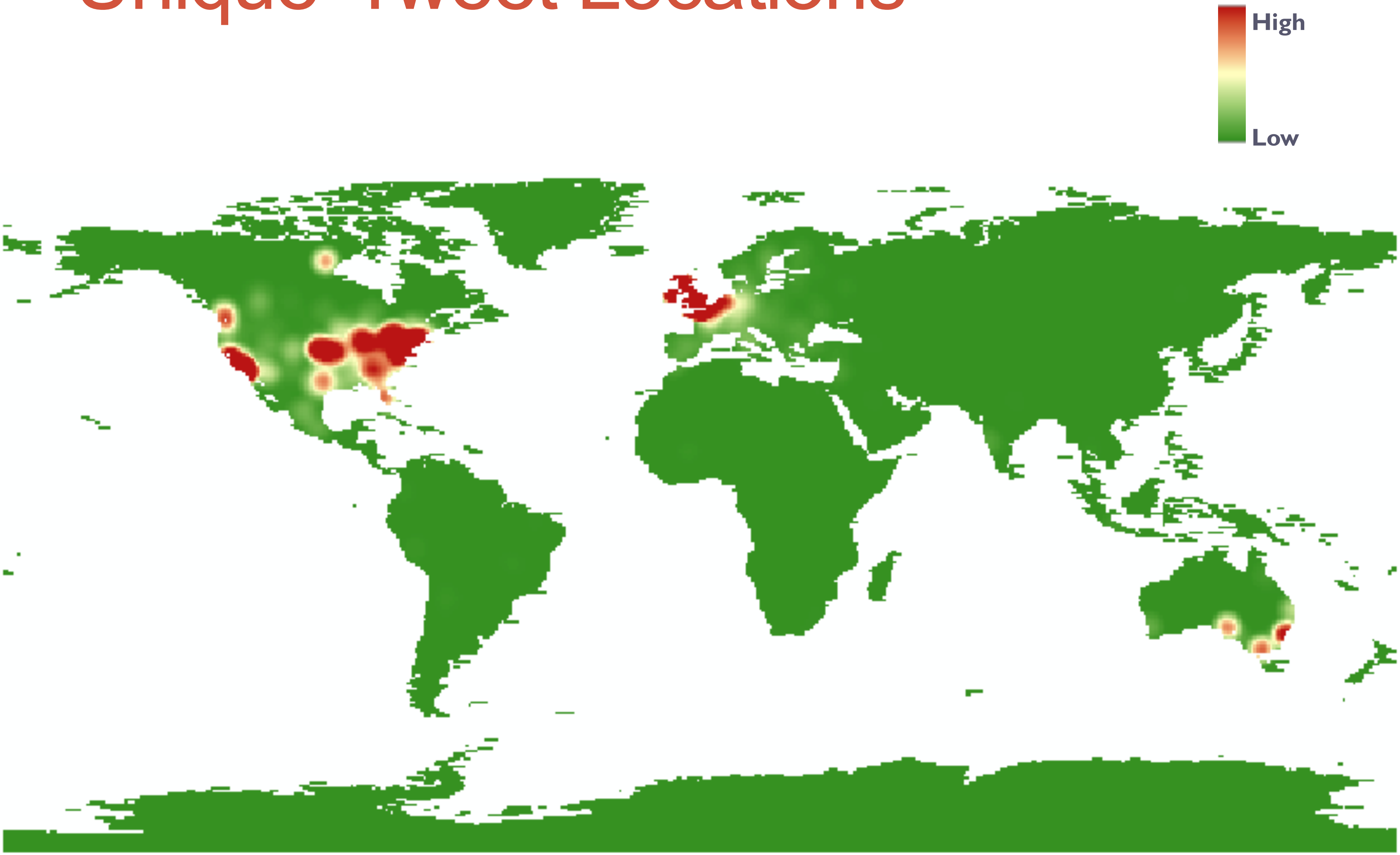
Ambiguous Place Names

- Single locations with multiple names:
  - Use ‘standard’ administrative data.
  - Deal only with the **coordinates** associated with each location from the geocoder.
  - If administrative information is required, it should then be extracted to the tweets using the coordinates.

Ambiguous Place Names

- Multiple locations with the same name:
  - **Tobler’s law** (*Everything is related to everything else, but near things are more related than distant things*).
  - Locations based other (**non-ambiguous**) tweets collected on the same topic.
  - Rankings determined by the **density** of non-ambiguous tweets at each ambiguous location.

‘Unique’ Tweet Locations

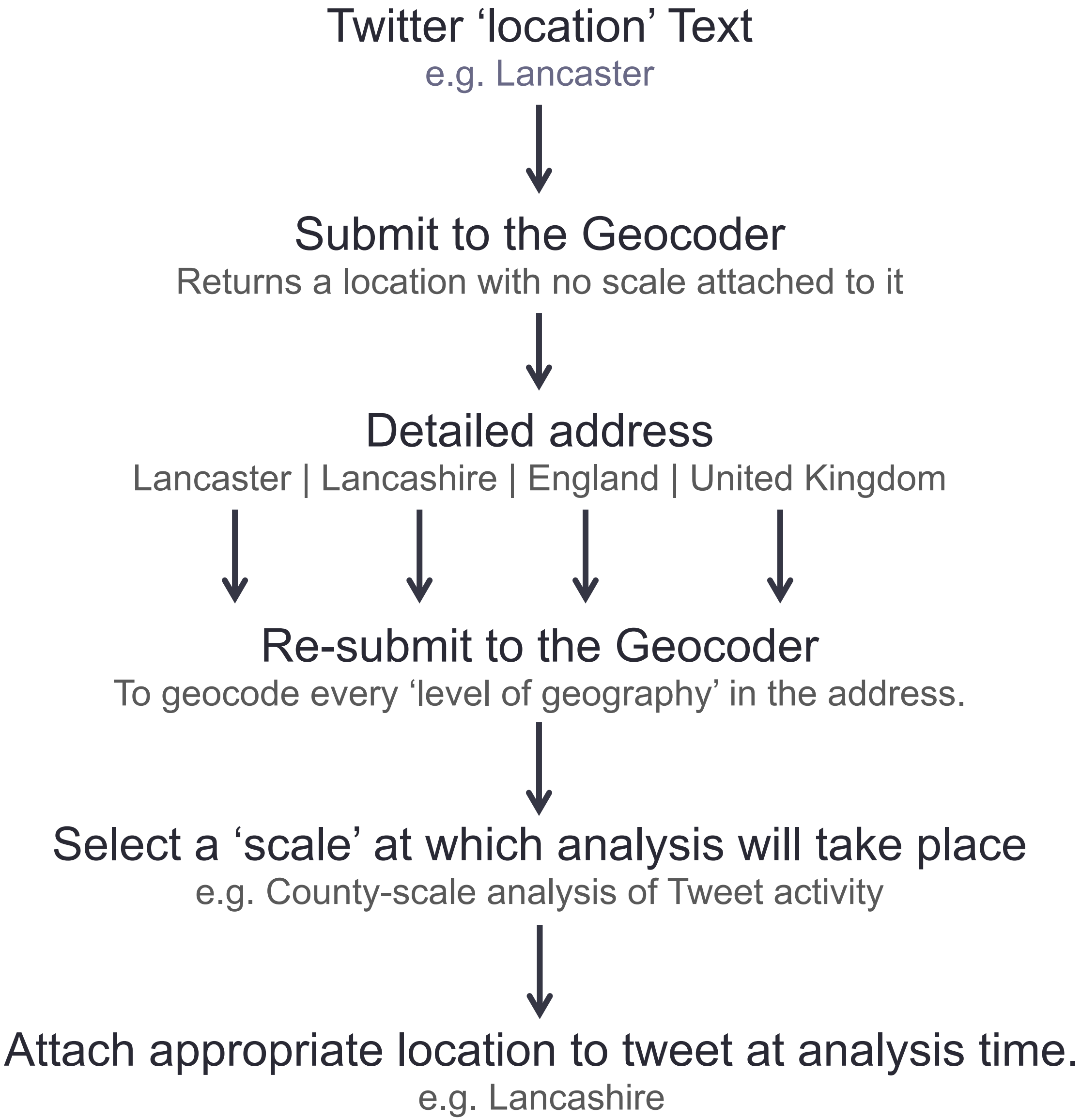


Methodology 2:

Undefined Level of Detail

Undefined Level of Detail

- The aim is to **standardise** the level of detail
  - Retrieve all of the **address components** for each tweet with the geocoder.
  - Get **coordinates** for each address component individually.
- At analysis time, locations of the required level of detail are used.
- Data with locations at insufficient detail are **discarded** from the analysis.



Twitter ‘location’ Text

e.g. Lancaster

Submit to the Geocoder

Returns a location with no scale attached to it

Detailed address

Lancaster | Lancashire | England | United Kingdom

Re-submit to the Geocoder

To geocode every ‘level of geography’ in the address.

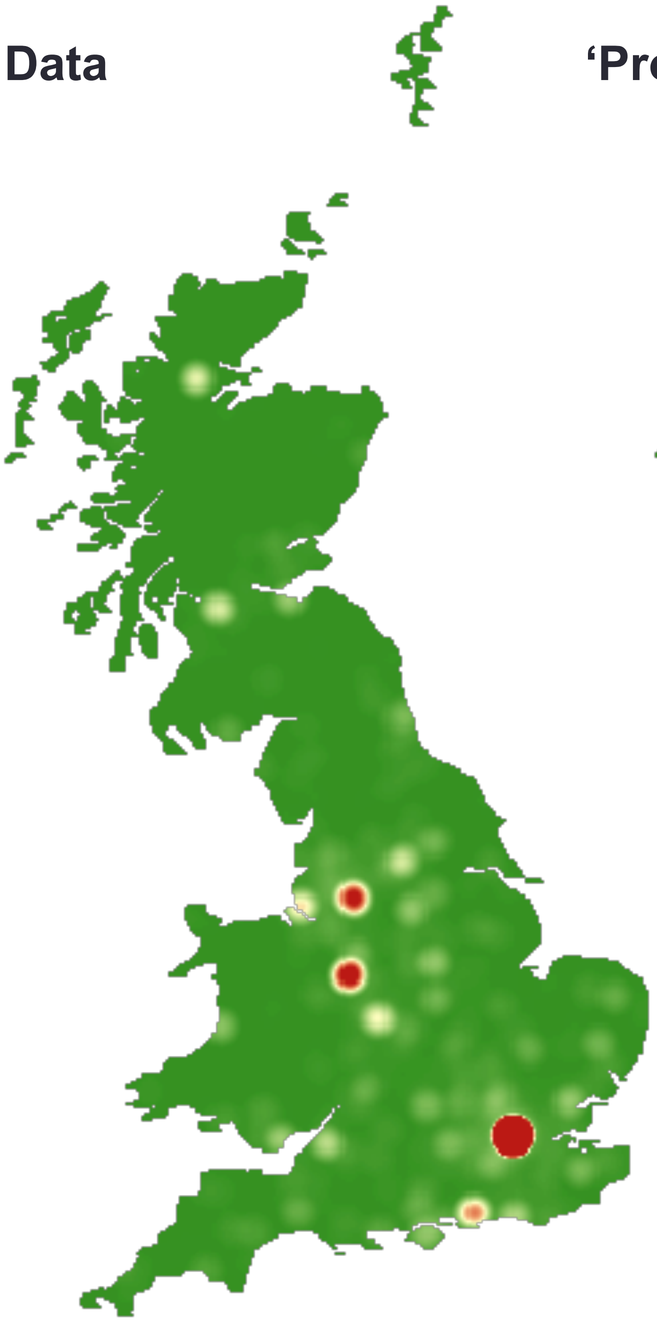
Select a ‘scale’ at which analysis will take place

e.g. County-scale analysis of Tweet activity

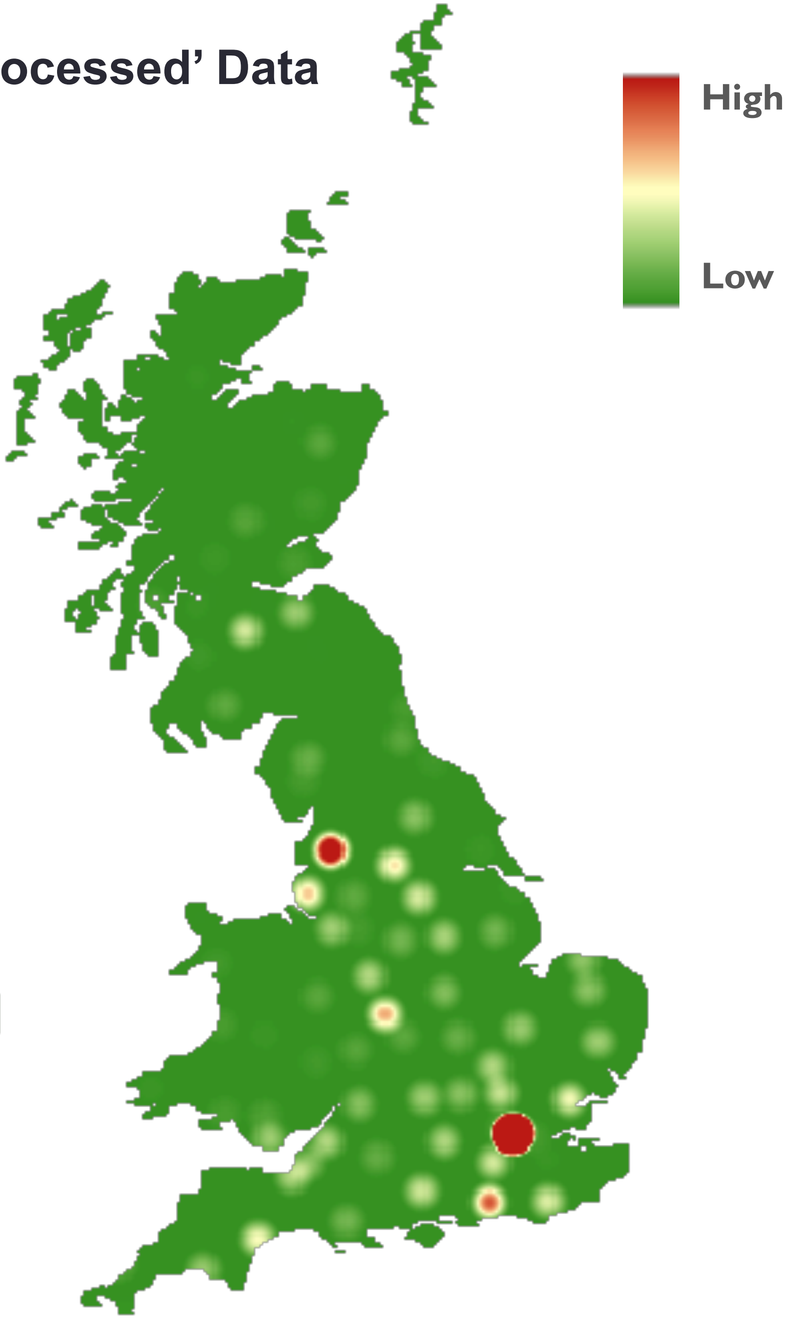
Attach appropriate location to tweet at analysis time.

e.g. Lancashire

‘Raw’ Data



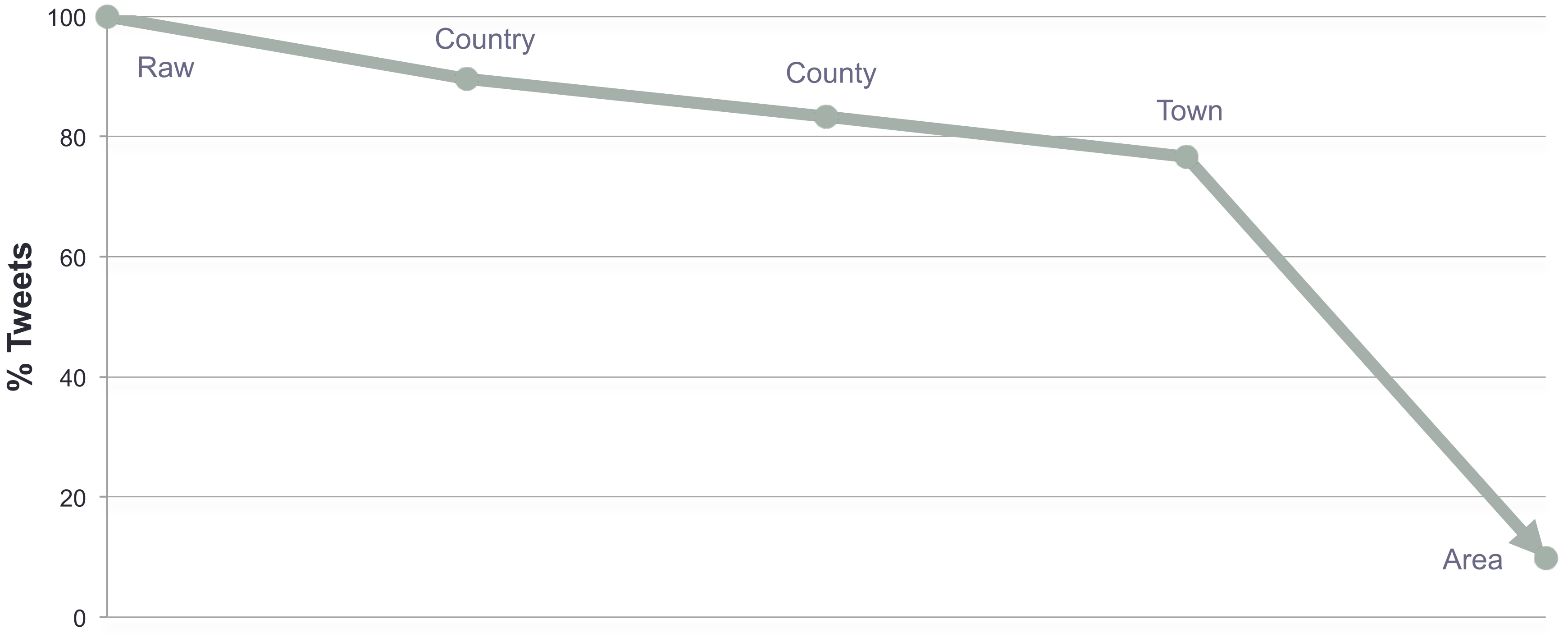
‘Processed’ Data



Data at an Undefined Scale

- ‘Trade-off’ :

scale of analysis vs ***volume of data***



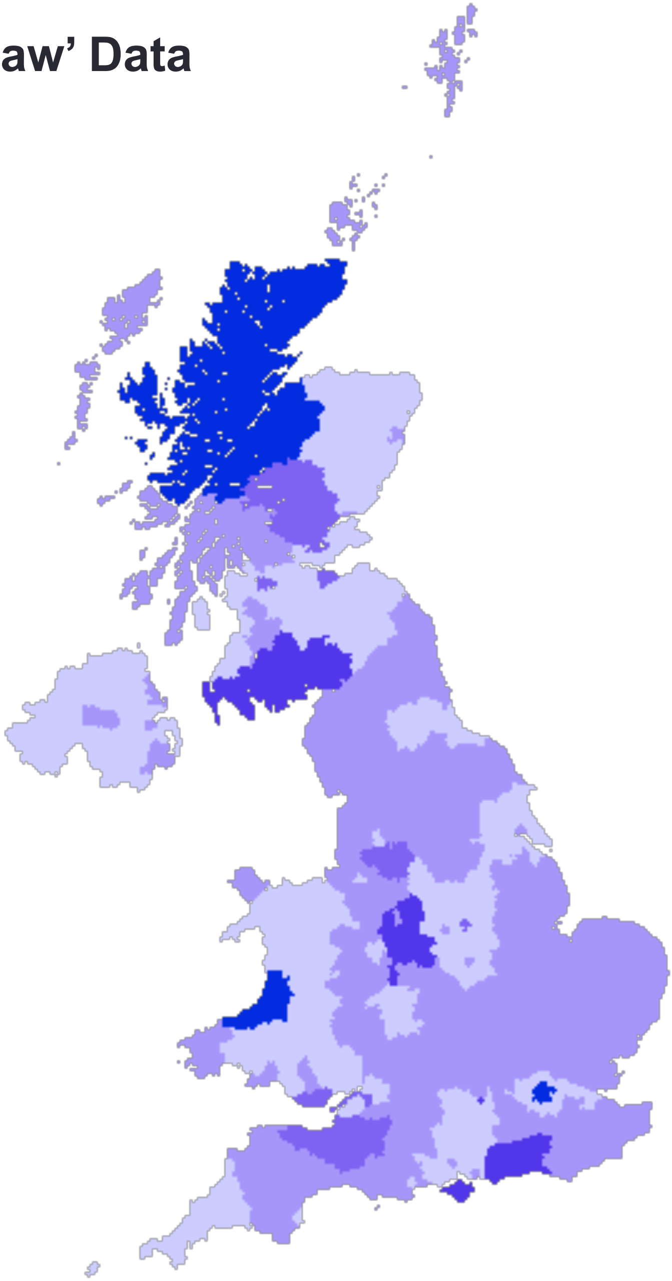
Why Does This Matter?

Case Study

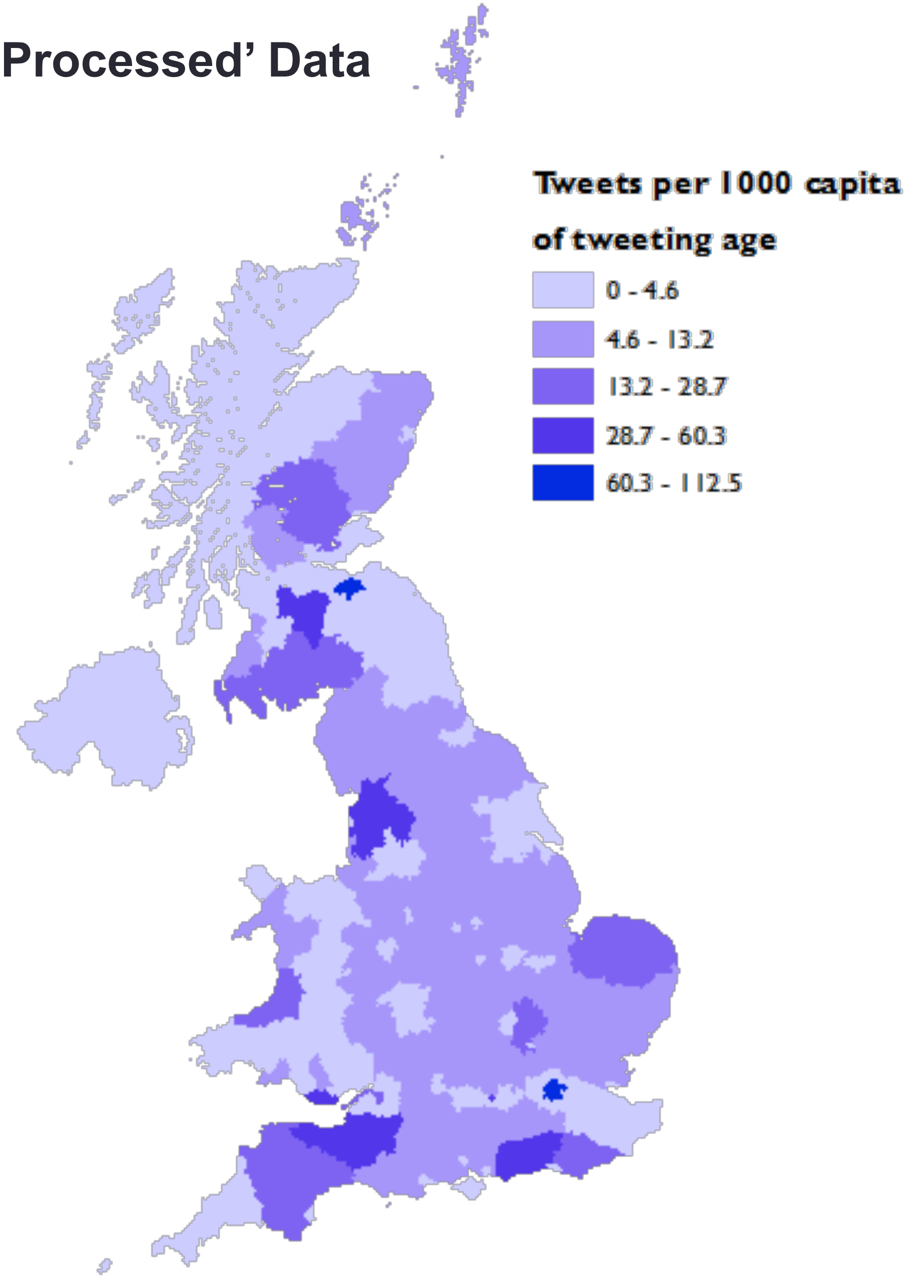
Why Does this Matter?

- Number of tweets per 1000 people of tweeting age, across the Counties and Unitary Authorities in the UK.
- Tweeting age was determined as being 10-59.
- A count of tweets was taken for each county, and normalised for the 'tweeting population' in that county.

‘Raw’ Data



‘Processed’ Data



Summary

- Data derived from social websites are frequently and increasingly used in spatial analysis.
- The locations attached to such data tend to rely on **place names**:
  - **non-unique**
  - lacking information regarding **level of detail**.

Summary

- This poses two issues in attempting to geocode the data:
    - Establishing which ‘place’ is the correct one;
    - The introduction of **false hotspots**.
- A methodology is demonstrated to address these issues:

Summary

- It has been demonstrated that this methodology has a significant impact upon analysis of this data.
- Our example was very **UK-Centric** , but these issues have a global significance, and are **intensified at the global scale**.
- Geocoders are powerful, but can be misleading if taken at face value.

Questions?