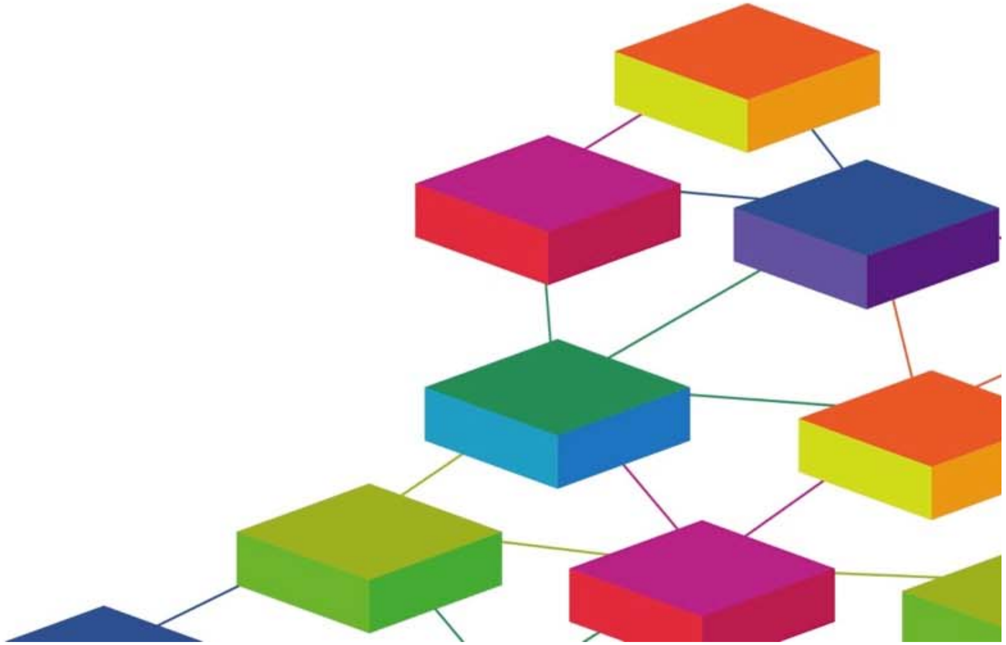
Consumer Data Research Centre

An ESRC Data Investment

The Provenance of Consumer and Social Media Data

Paul Longley
and colleagues



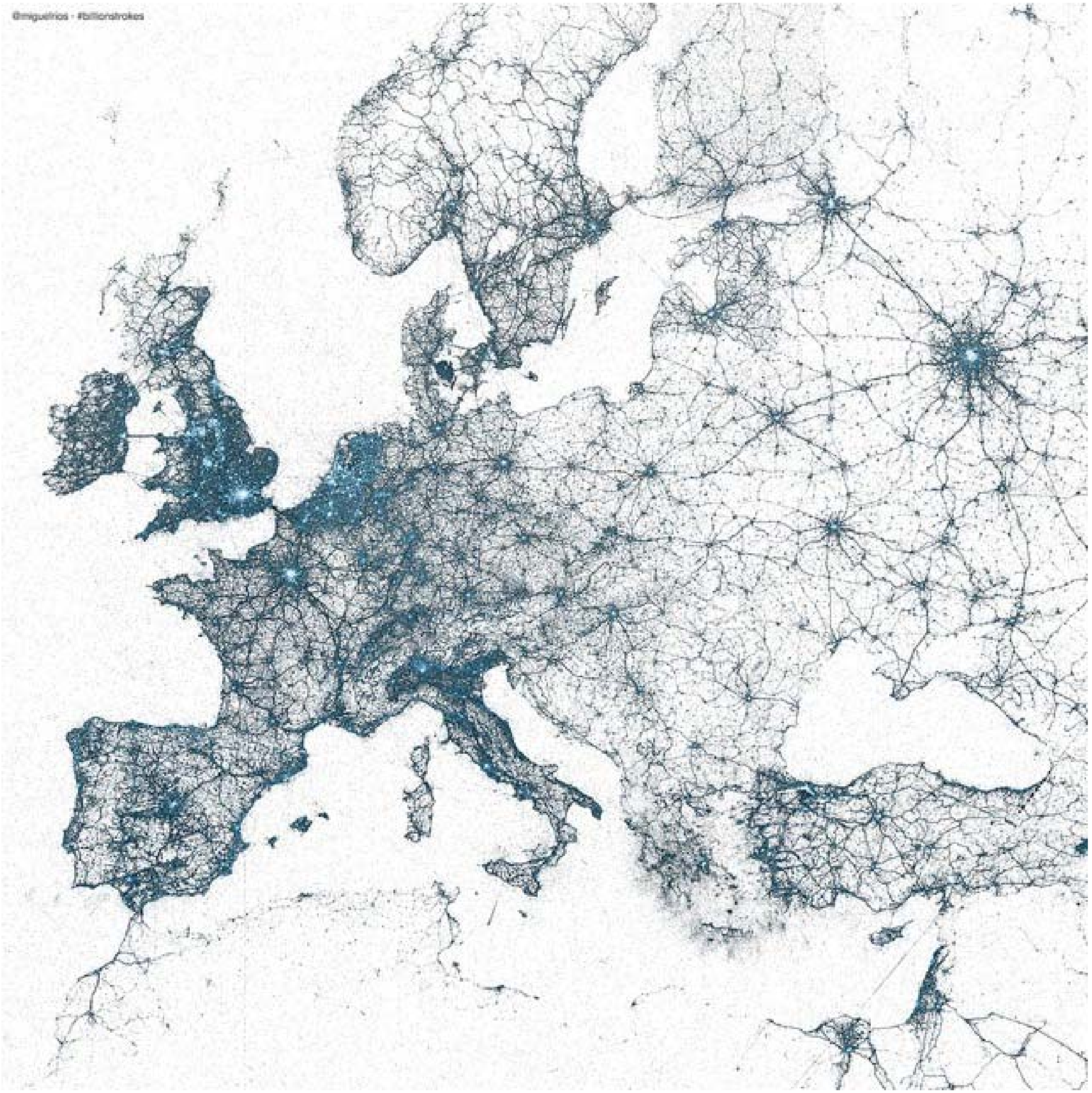
4th International Conference on Data Management Technologies and Applications, Colmar, 20-22 July

‘Big’ consumer data

- Real share of data: ‘exhaust’
- Naïve empiricism?
- SDI?
  - Incompatible measures and units
  - Big (geotemporal) Data and the linear research design
  - Data linkage (but to which ‘populations’?)
- Front loading of modelling assumptions to model individuals through space and time
  - Horses for courses approach to data creation and maintenance

Tweets – pretty but what value?

The European distribution of a billion global Tweets between 2011 and mid-2013.

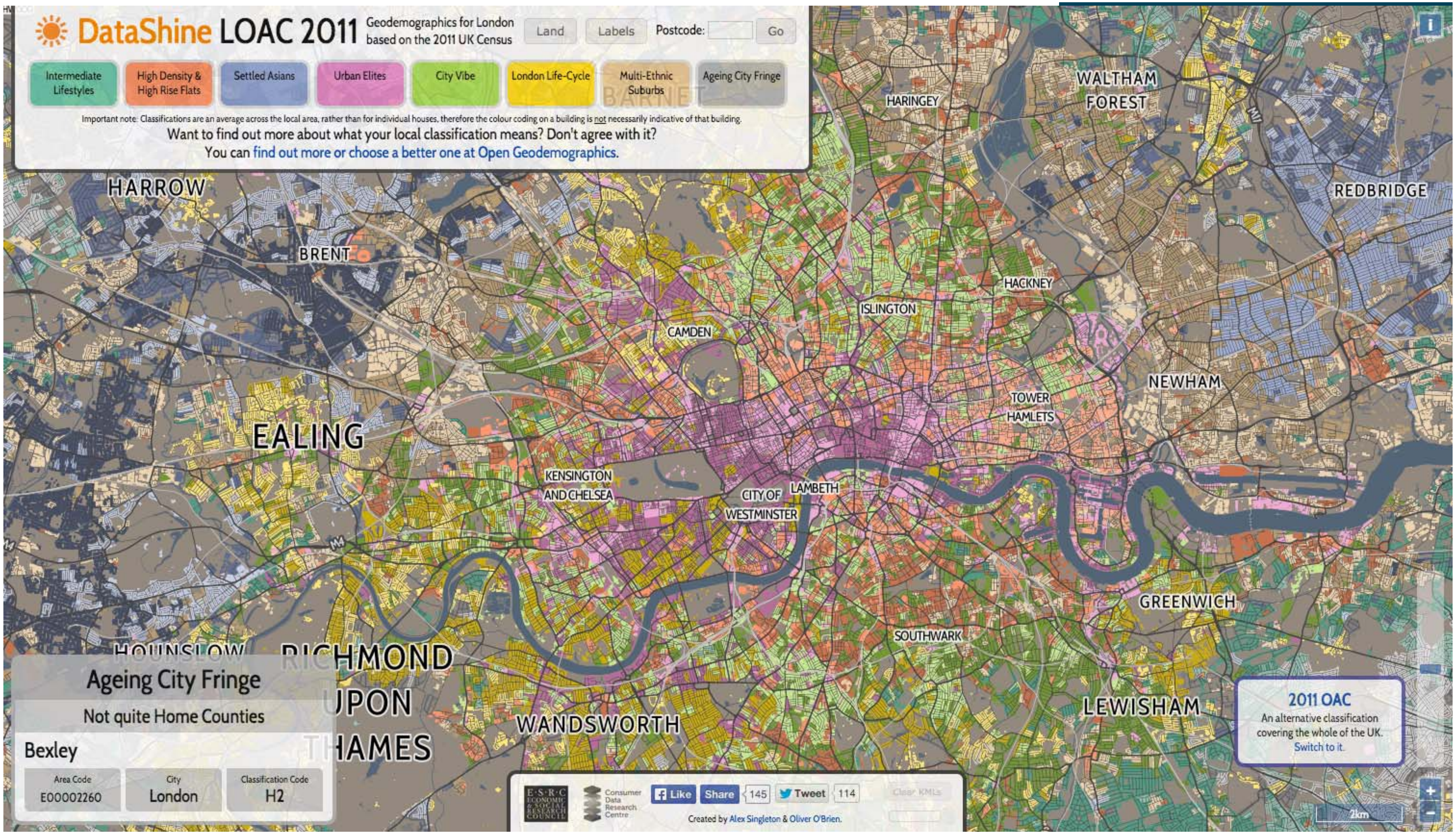


‘Big’ consumer data

- Real share of data: ‘exhaust’
- Naïve empiricism?
- SDI?
  - Incompatible measures and units
  - Big (geotemporal) Data and the linear research design
  - Data linkage (but to which ‘populations’?)
- Front loading of modelling assumptions to model individuals through space and time
  - Horses for courses approach to data creation and maintenance

Consumer Data Research Centre (CDRC)

- Multi- institution laboratory (c.£12m) that discovers, mines, analyses and synthesises consumer-related datasets from around the UK.
- Creates, supplies, maintains and delivers consumer-related data to a range of end users
  - CDRC-Public (Open, maps)
  - CDRC-Stakeholder / Archive
  - CDRC-Secure
- Programme of research and outreach activities.



Consumer Data Research Centre

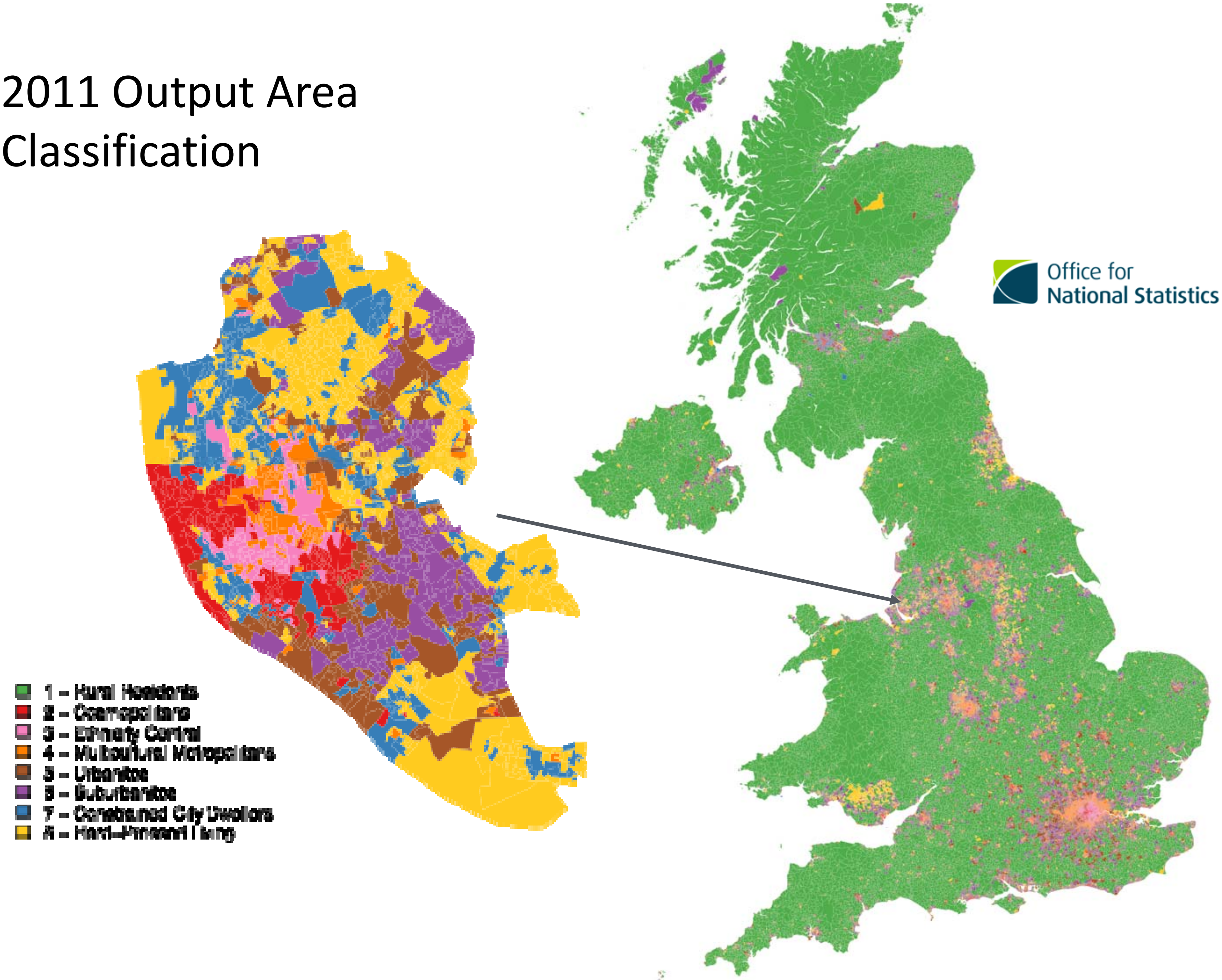
An ESRC Data Investment

London Output Area Classification

Geo-temporal demographics

- Kaleidoscope and mosaic
- Activities not night time residence (Alex Singleton)
- Process and dynamics known, but not generalised pattern that they fit

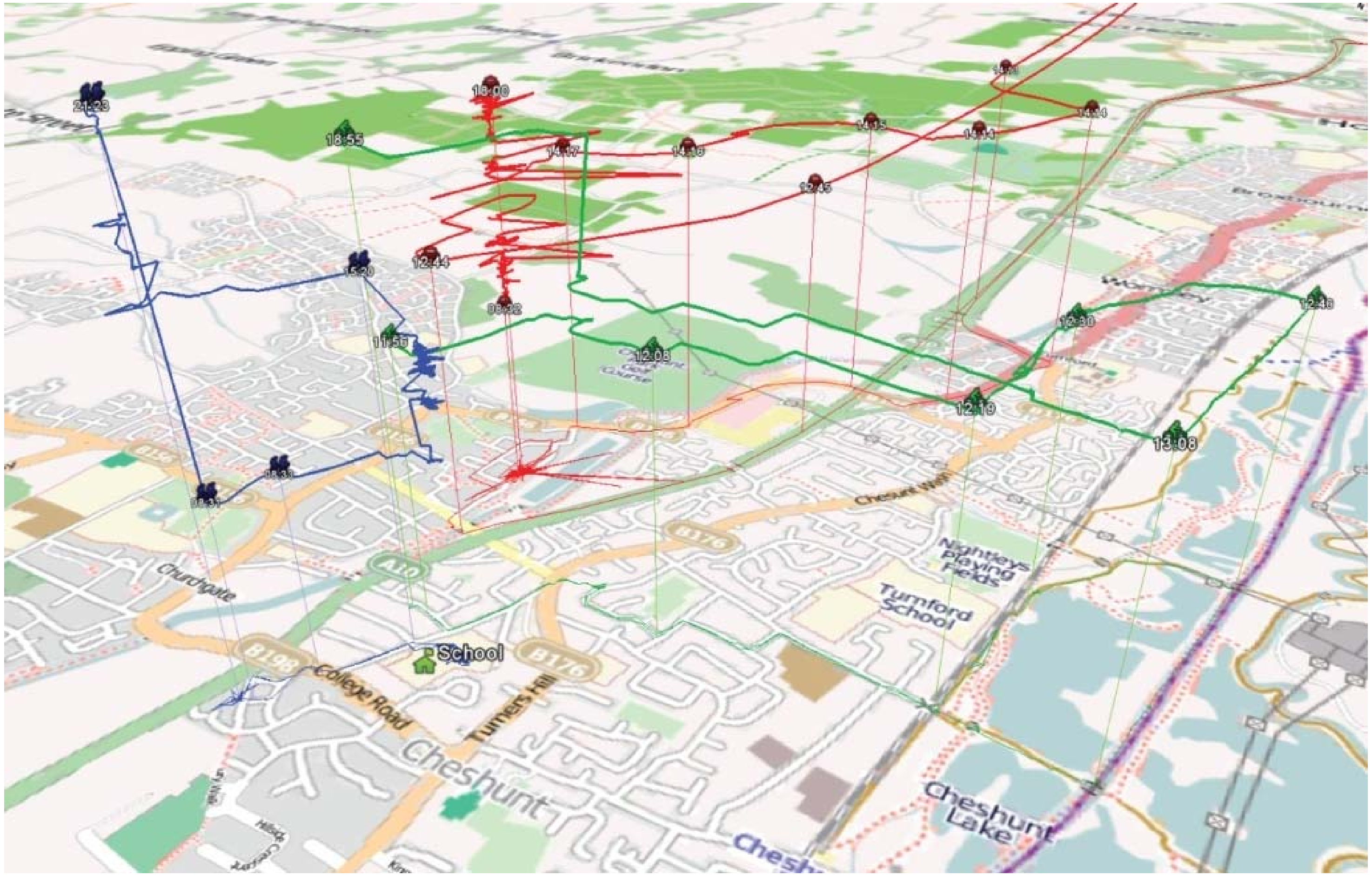
2011 Output Area Classification




http://www.google.co.uk/intl/en_uk/earth/
52: POORER FAMILIES, MANY CHILDREN, TERRACED HOUSING
51: YOUNG PEOPLE IN SMALL, LOW COST TERRACES
Urban Adversity
Affluent Achievers
11: SETTLED SUBURBIA, OLDER PEOPLE
59: DEPRIVED AREAS AND HIGH-RISE FLATS

Geo-temporal demographics

- Kaleidoscope, not mosaic
- Activities not night time residence (Alex Singleton)
- Process and dynamics known, but not generalised pattern that they fit



Schematic representation of weekend activities of three children in Cheshunt, UK.

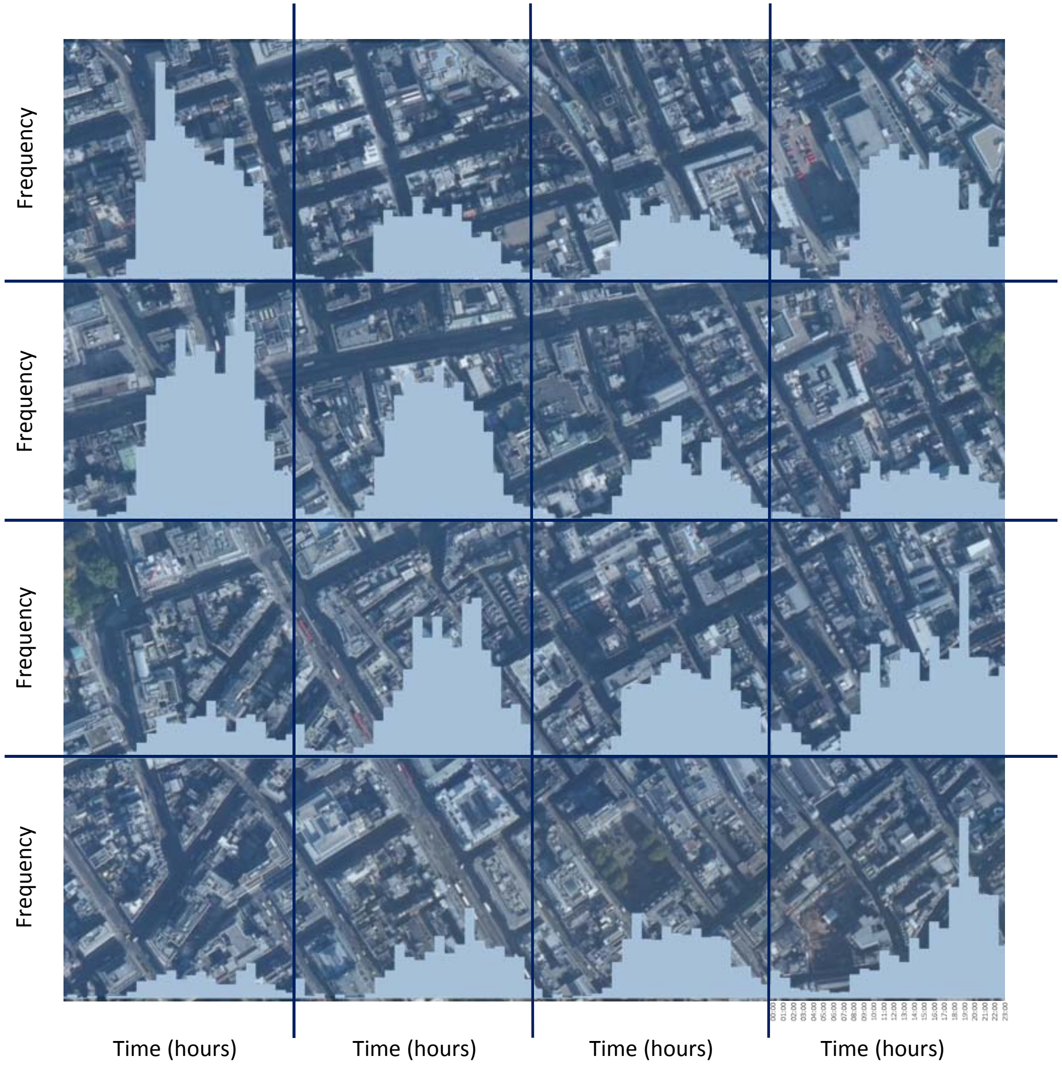
(Reproduced with permission of Yi Gong: base image Courtesy www.openstreetmap.org)

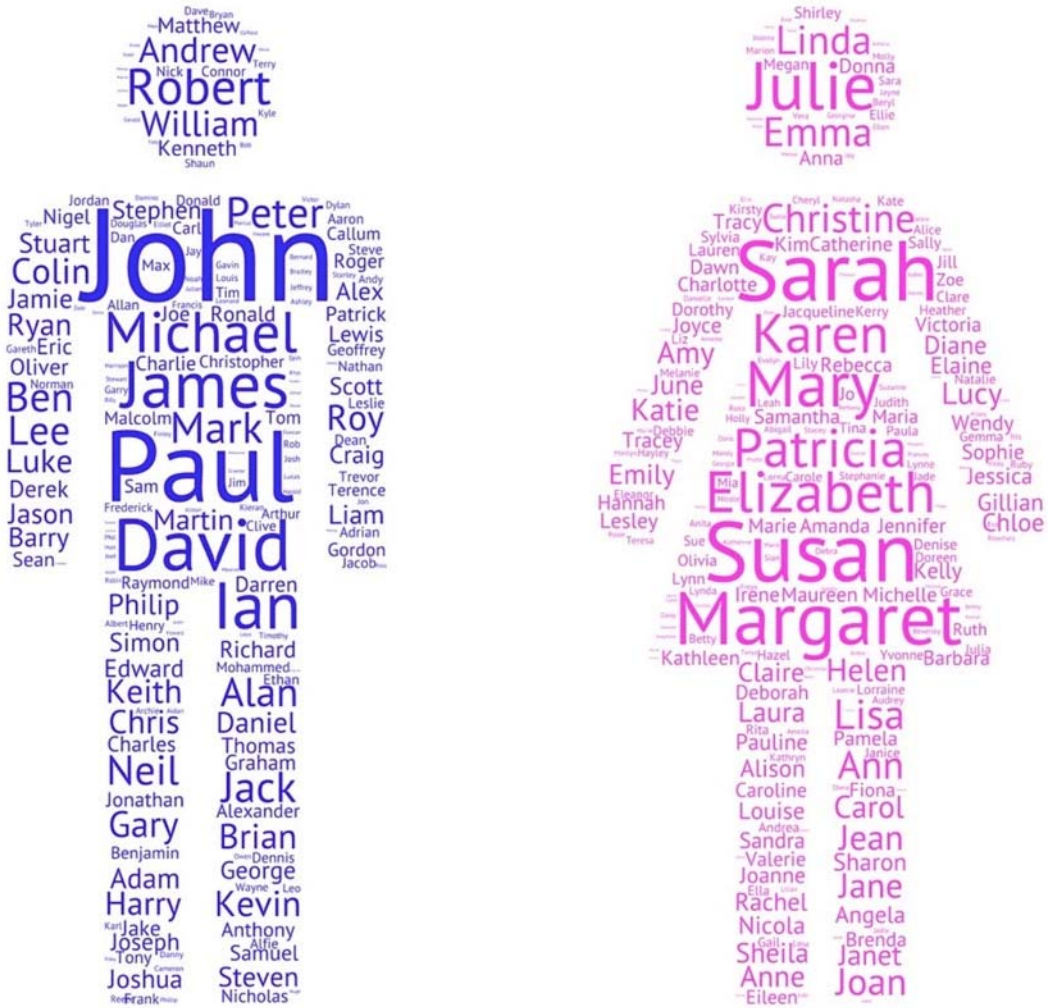
Case Study (1): Twitter demographics

Twitter estimated footfall in Soho

- The frequency of geotagged Tweets across space and time can tell us about the dynamics of a city (courtesy Guy Lansley)

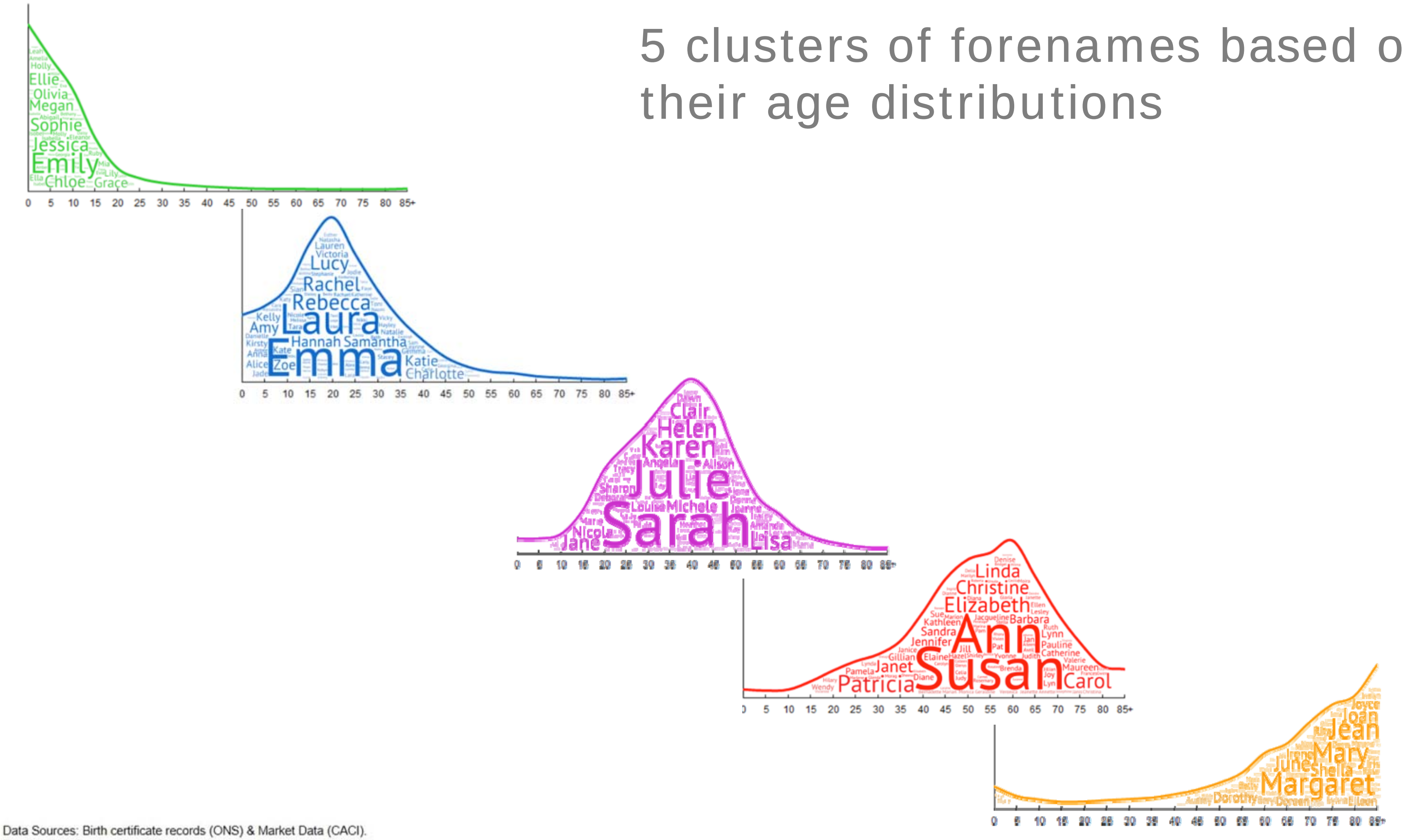
The average weekday activity in 2013





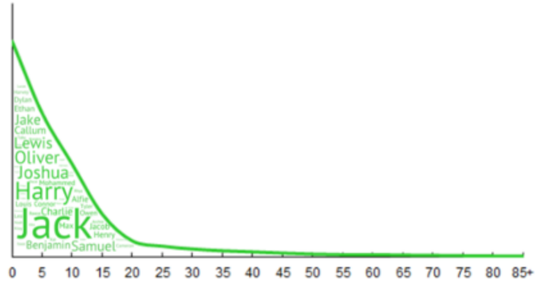
Data Sources: Birth certificate records (ONS) & Market Data (CACI).
Produced by Guy Lansley, UCL

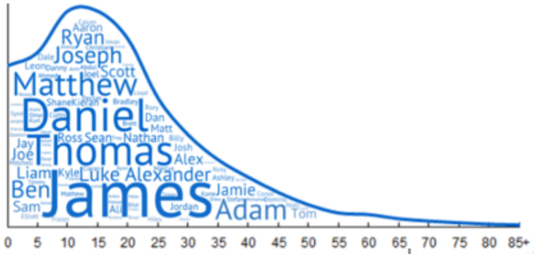
5 clusters of forenames based on their age distributions

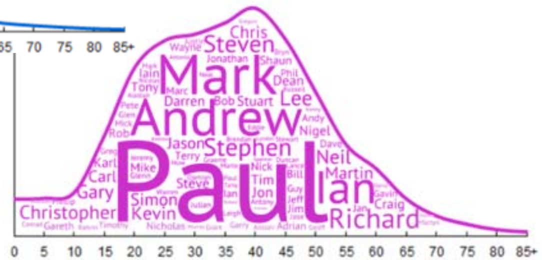


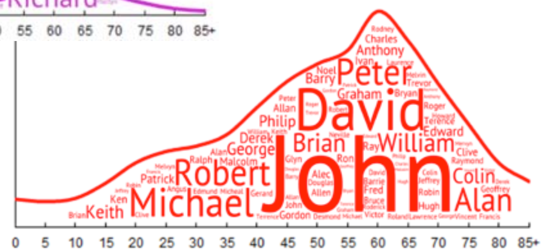
Data Sources: Birth certificate records (ONS) & Market Data (CACI).
Produced by Guy Lansley, UCL

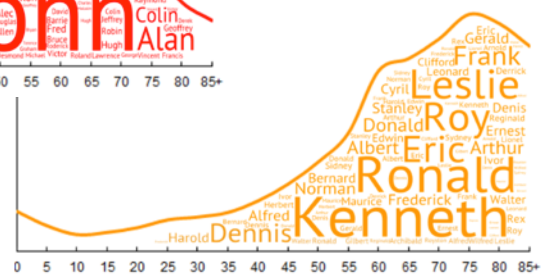
5 clusters of forenames based on their age distributions











Data Sources: Birth certificate records (ONS) & Market Data (CACI).
Produced by Guy Lansley, UCL

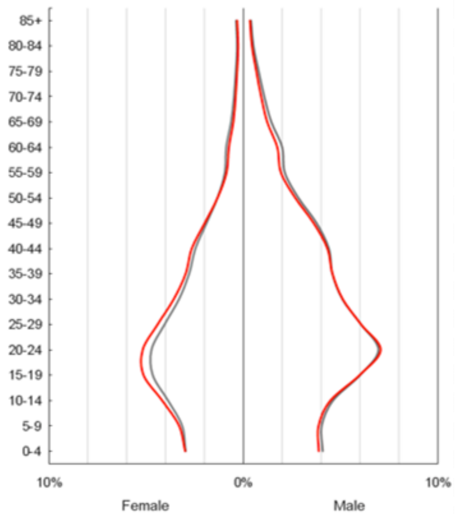
Inferred demographic structure of Tweeters

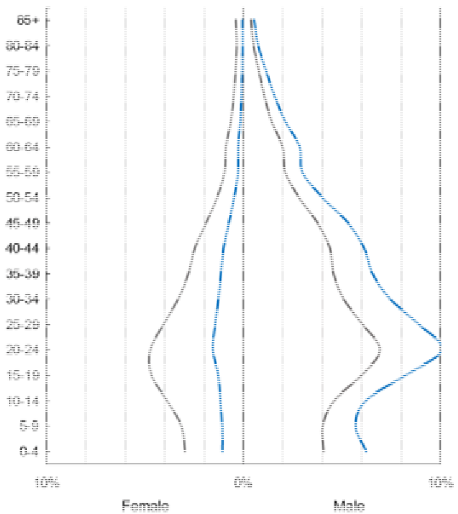
The O2 Arena

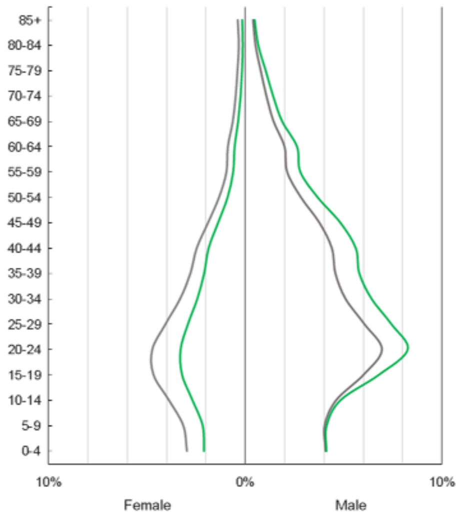
The Emirates stadium

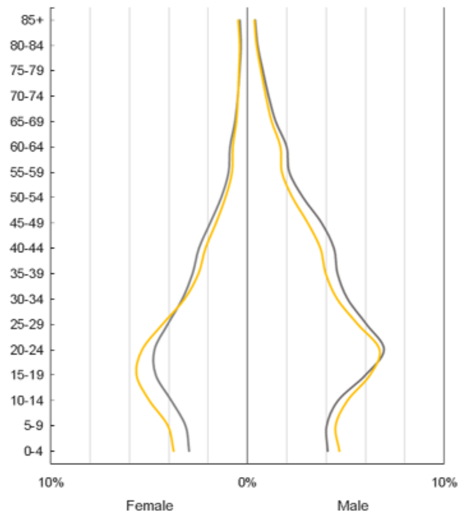
Canary Wharf

Westfield Stratford







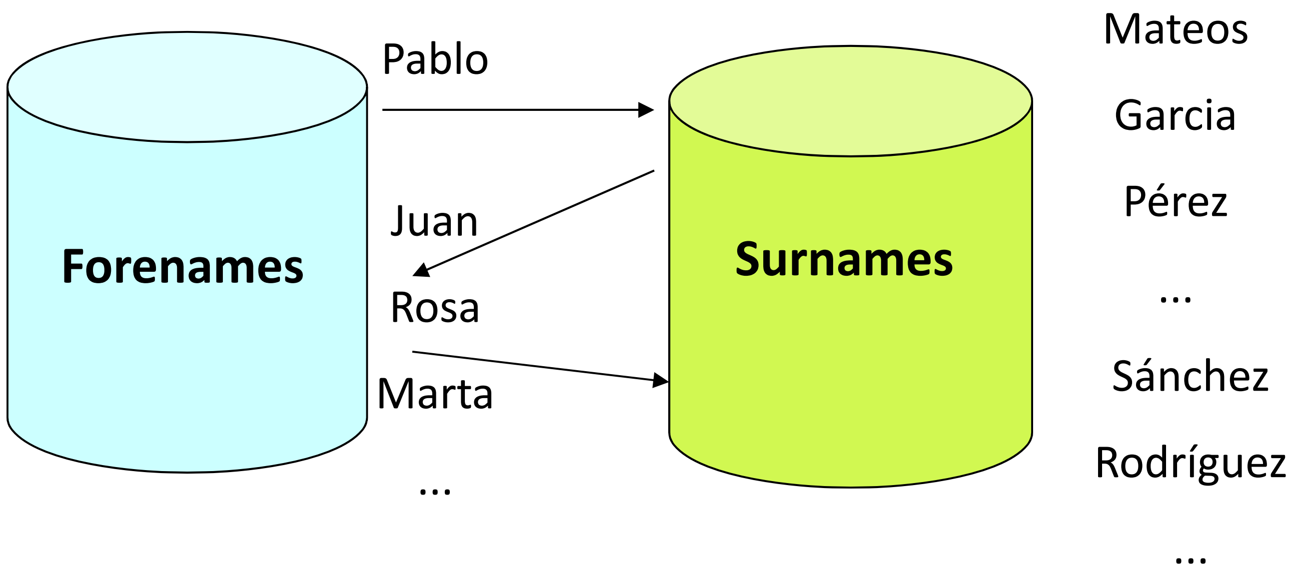


Onomap classification

Forename-Surname clustering

(based on Hanks and Tucker, 2000)

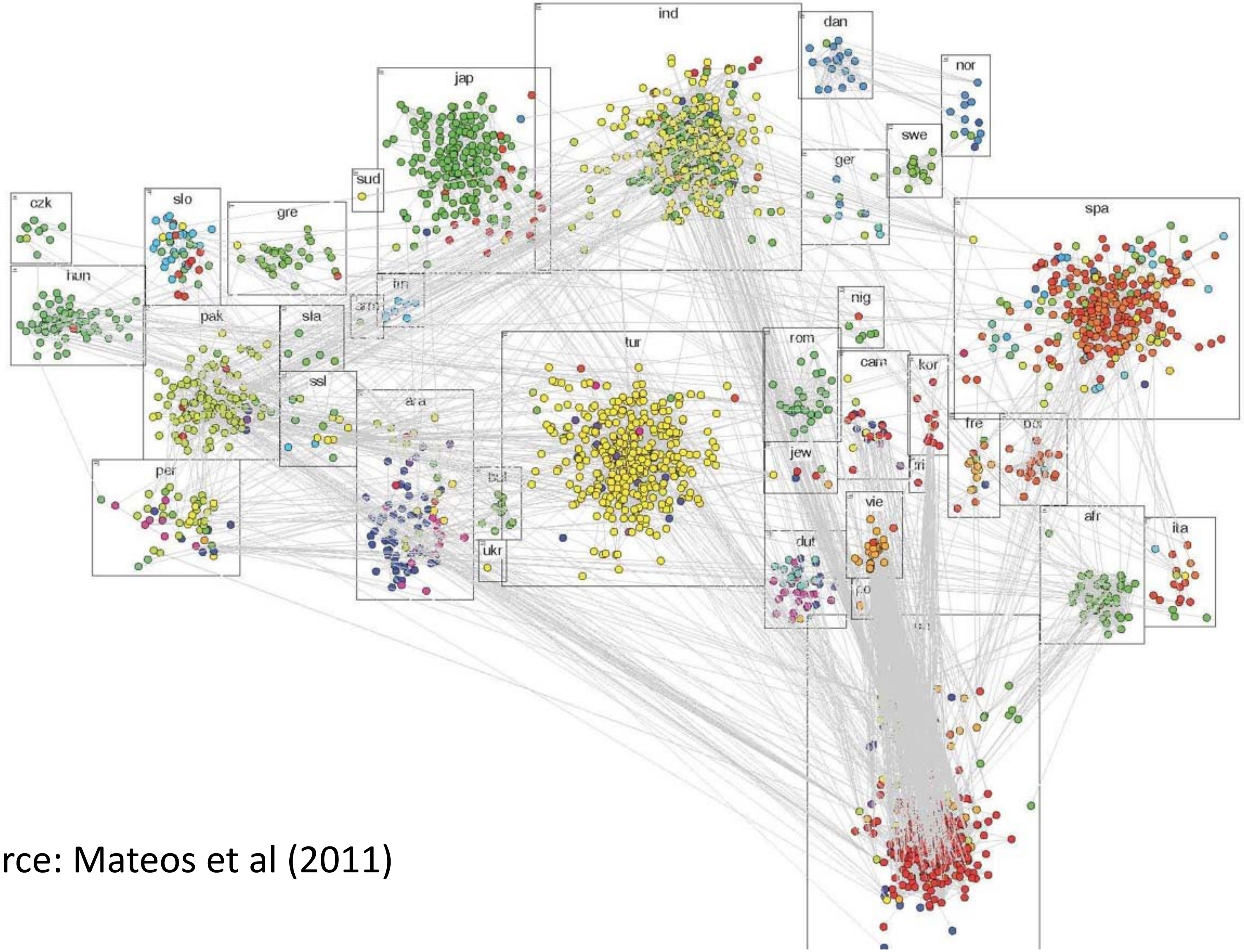
Global names registers



- Several iterations until self-contained cluster is exhausted
- Cluster assigned a cultural, ethnic & linguistic Onomap type
- Probability of ethnicity assigned to each name

Mateos et al (2007) CASA Working Paper 116

WorldNames CEL clusters

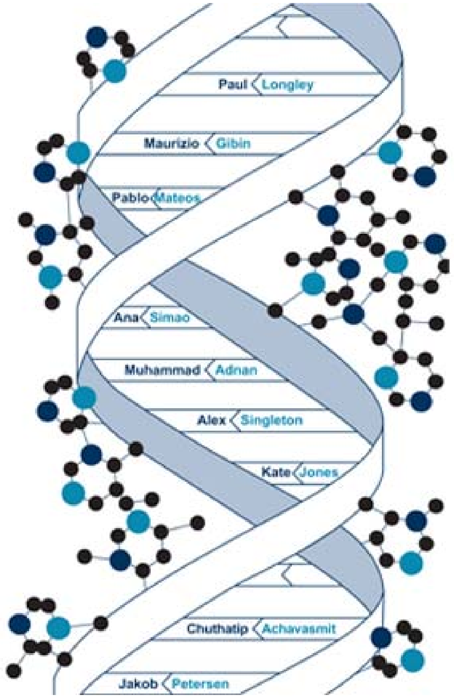


Source: Mateos et al (2011)

Cultural, Ethnic and Linguistic roots of names

OnoMAP is a new way of classifying people and the places they live, based on our common cultural, ethnic and linguistic roots.

OnoMAP analyses common patterns of forenames and surnames using one of the world's largest databases of people drawn from 28 countries. The OnoMAP classification covers over 500,000 forenames and 1 million surnames, and most exhibit distinctive geographic patterning.



ONOMAP

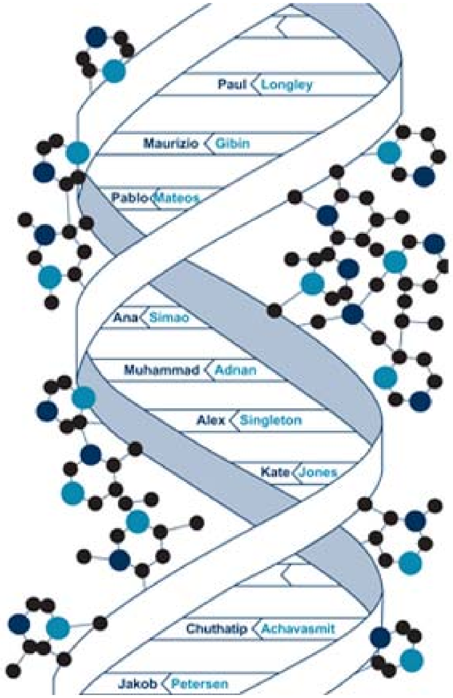
Forename
Surname
Search

Guy Lansley – English
Alyson Lloyd – Welsh
Kira Kowalski - Polish
Wen Li – Chinese
Jens Kandt – Danish
Muhammad Adnan – Pakistani
Syed Uddin - Bangladeshi

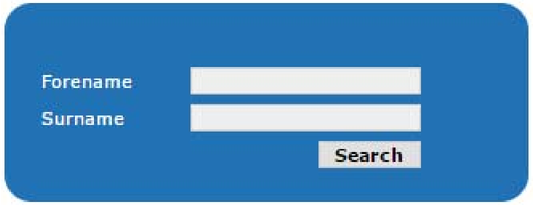
Cultural, Ethnic and Linguistic roots of names

OnoMAP is a new way of classifying people and the places they live, based on our common cultural, ethnic and linguistic roots.

OnoMAP analyses common patterns of forenames and surnames using one of the world's largest databases of people drawn from 28 countries. The OnoMAP classification covers over 500,000 forenames and 1 million surnames, and most exhibit distinctive geographic patterning.



ONOMAP



Guy Lansley – English
Alyson Lloyd – Welsh
Kira Kowalski - Polish
Wen Li – Chinese
Jens Kandt – Danish
Muhammad Adnan – Pakistani
Syed Uddin - Bangladeshi

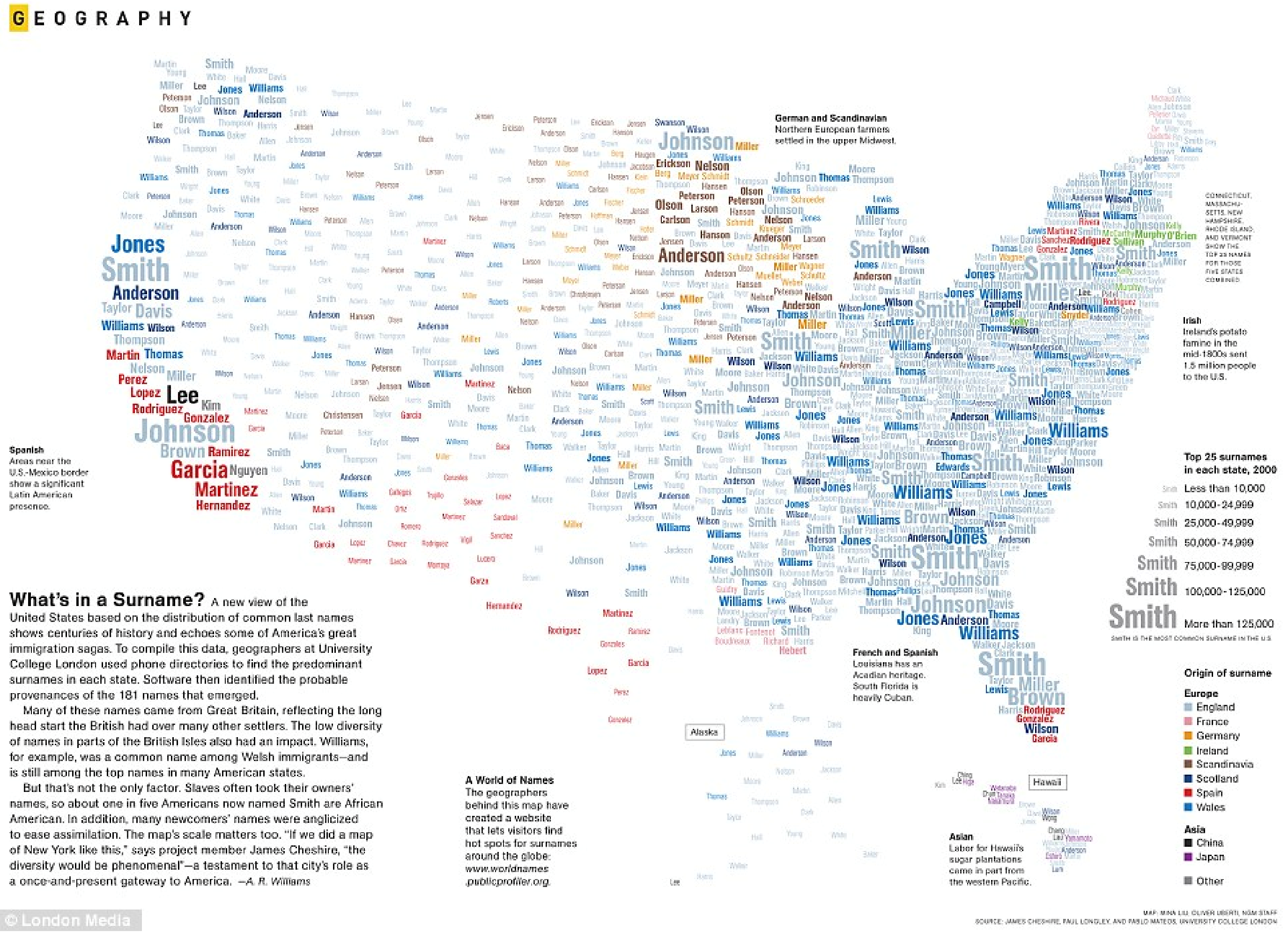
Consumer Data Research Centre

An ESRC Data Investment

4th International Conference on Data Management Technologies and Applications, Colmar, 20-22 July

UCL

What's in a Surname (NGM, 2011)



What's in a Surname? A new view of the United States based on the distribution of common last names shows centuries of history and echoes some of America's great immigration sagas. To compile this data, geographers at University College London used phone directories to find the predominant surnames in each state. Software then identified the probable provenances of the 181 names that emerged.

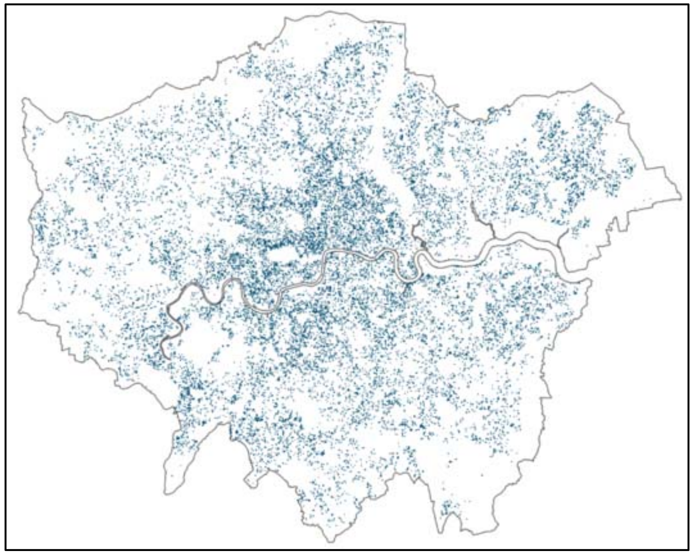
Many of these names came from Great Britain, reflecting the long head start the British had over many other settlers. The low diversity of names in parts of the British Isles also had an impact. Williams, for example, was a common name among Welsh immigrants—and is still among the top names in many American states.

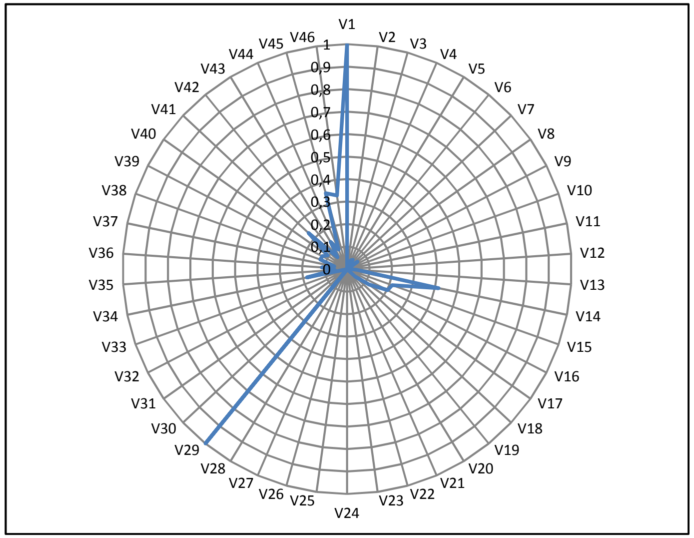
But that's not the only factor. Slaves often took their owners' names, so about one in five Americans now named Smith are African American. In addition, many newcomers' names were anglicized to ease assimilation. The map's scale matters too. "If we did a map of New York like this," says project member James Cheshire, "the diversity would be phenomenal"—a testament to that city's role as a once-and-present gateway to America. —*A. R. Williams*

James Cheshire

Geo-temporal Demographics of Social Media

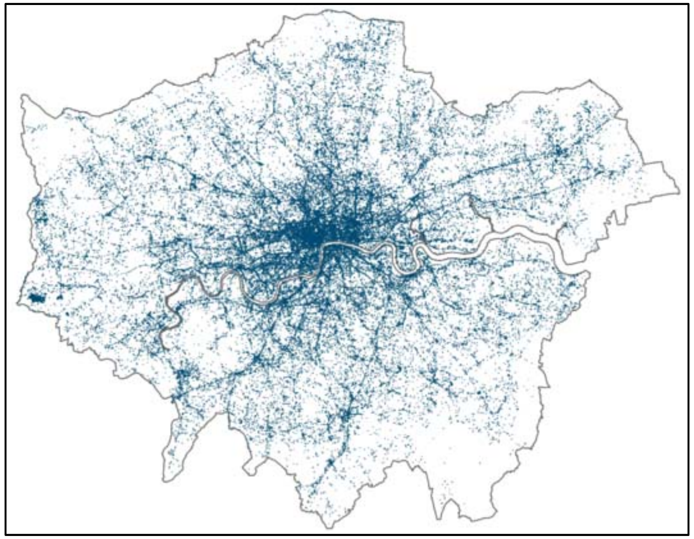
- **Group A: London Residents**
- **Tweets made near residential locations.**
- **Tweets made on weeknights or weekends.**

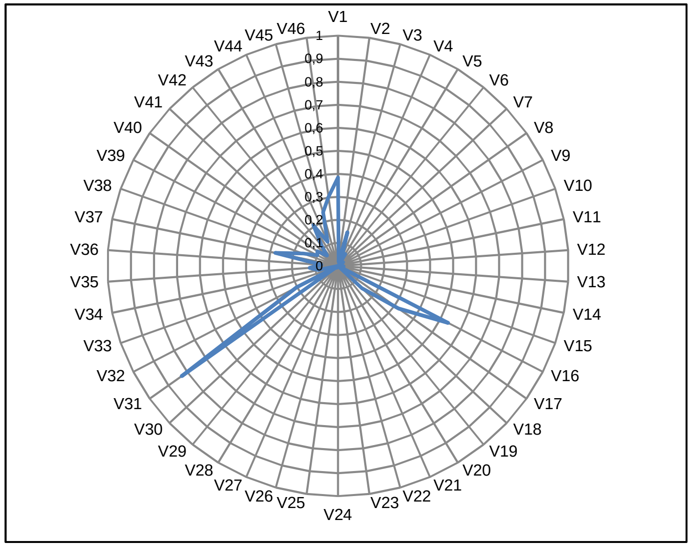




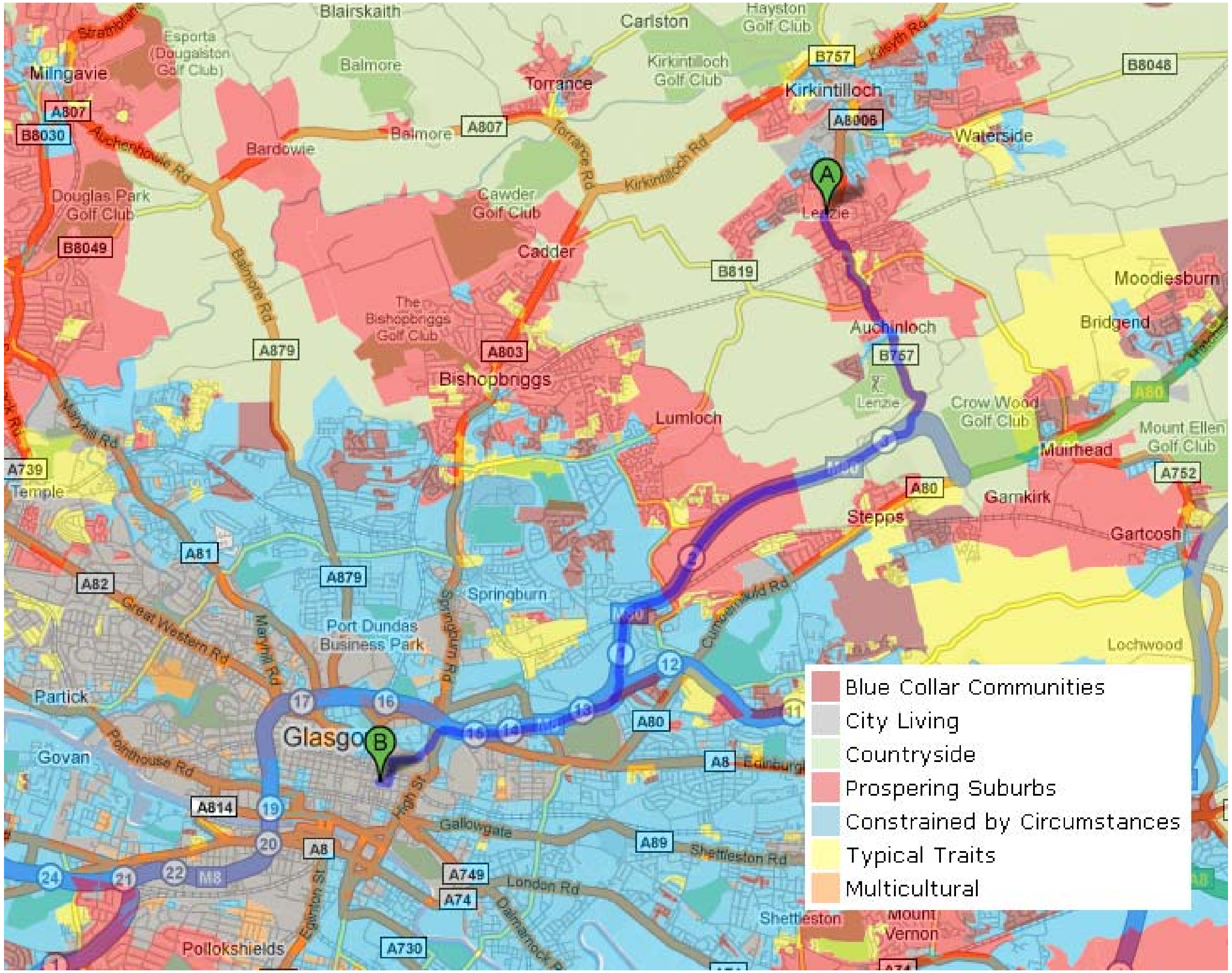
Geo-temporal Demographics of Social Media

- **Group D: The Daily Grind**
- **Tweets made during peak weekdays and nights.**
- **Sent from residential locations or in transit.**



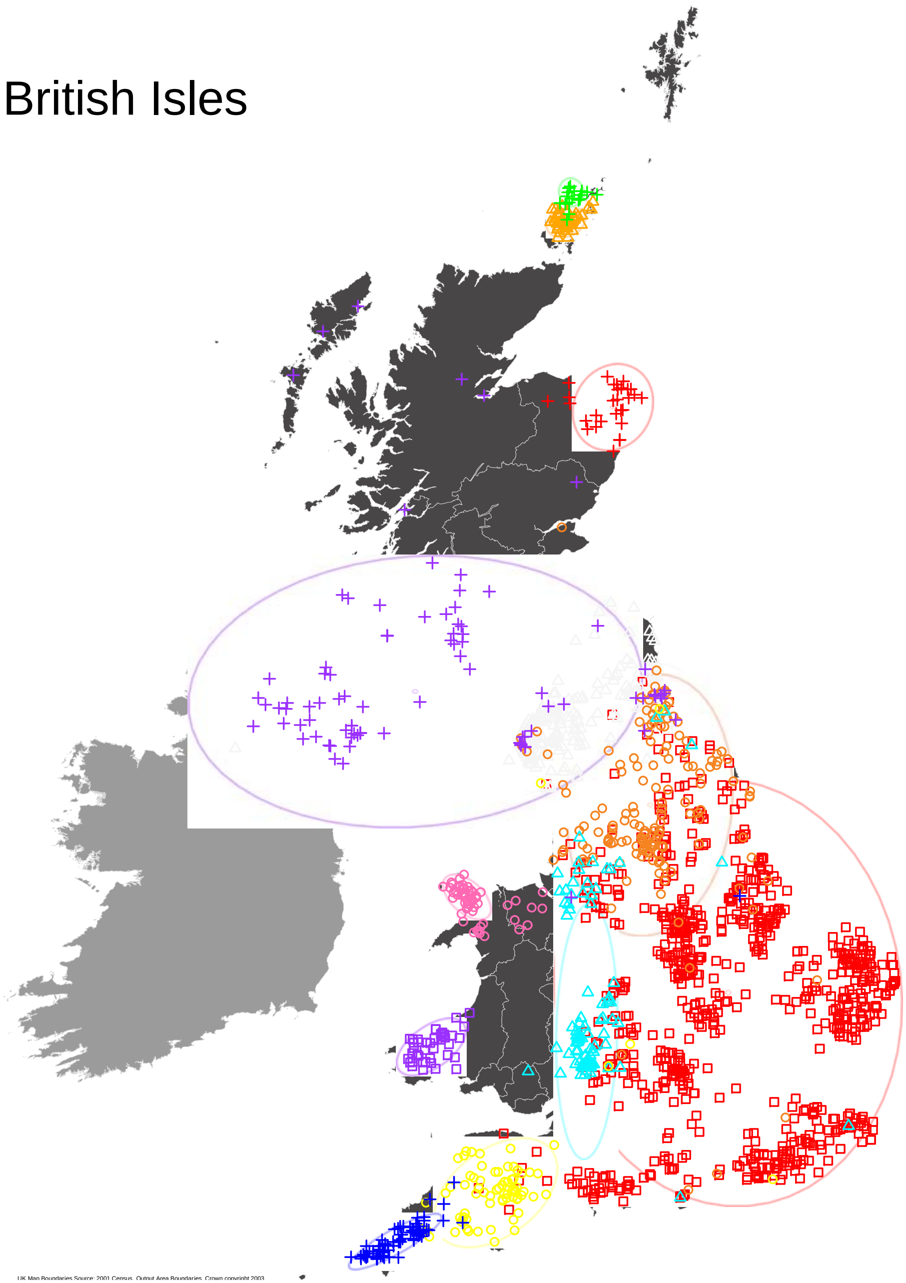


Case study (2): Social mobility and life chances



The local geodemography of Glasgow, showing the 7.8 mile route that links communities with life expectancies of 54 and 82

The People of the British Isles

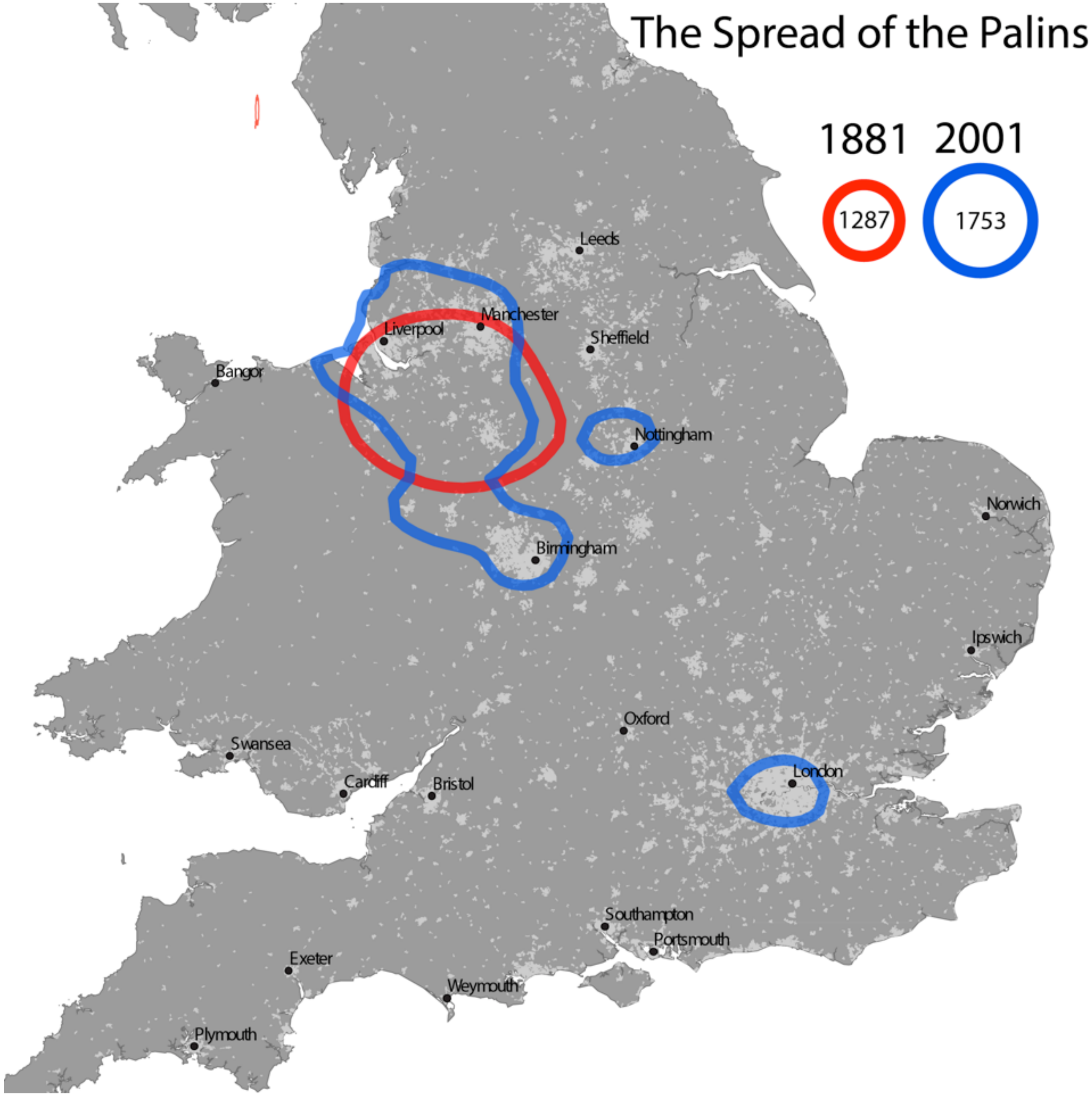


Names in Great Britain

TABLE 1: A CATEGORISATION OF BRITISH SURNAMES. ADAPTED FROM BARKER ET AL., 2007.

Category	Example	Explanation
Occupational (Metonyms)		
Profession	Smith	Blacksmith/ metal worker
Office/ Trade	Reeve	Chief magistrate/ overseer
Rank/Status	Knight	A knighted person
Occupation Features	Falconer	One who kept/trained Falcons
Local Surnames (50% of surnames)		
Toponymic (from landscape)	Rivers	Dweller near river
Toponymic (from village/ region)	Cornwall	Man from Cornwall
Habitation (residence)	Gate	Habitation at/near a gate
Habitation (work)	Hall	A worker at the hall.
Surnames of Relationship		
From personal name (patronymic)	Johnson/ Jones	Son of John
From personal name (metronymic)	Margaretson	Son of Margaret
Personal name from other relative	Also: Johnson	Related to John
Personal name from diminutive	Dickens	Son of Dick (Richard)
Clan or tribal names	MacBain	Related to the MacBain clan.
Nicknames		
From animals	Fox	Slyness or other attributes
From characteristic traits	Careless	Free from care/ responsibility
From objects	Shorthose	Someone who wore short boots
From physical features	Little	A small person
From times and seasons	Pasque	Person born at Easter
From iconic description	Drinkwater	Heavy drinker

Pablo Mateos, Barker et al 2007



Courtesy:
James
Cheshire

gbnames.publicprofiler.org

Social Demographics

Social Demographics	Statistics
Category of surname	Celtic; Irish; Starting with O-
Mosaic type with highest index #	Counter Cultural Mix
Index of top Mosaic type *	227
% of people with a more rural name	94
% of people with a more high-status name	92
Cultural, Ethnic, Linguistic categories of surname	British, Irish

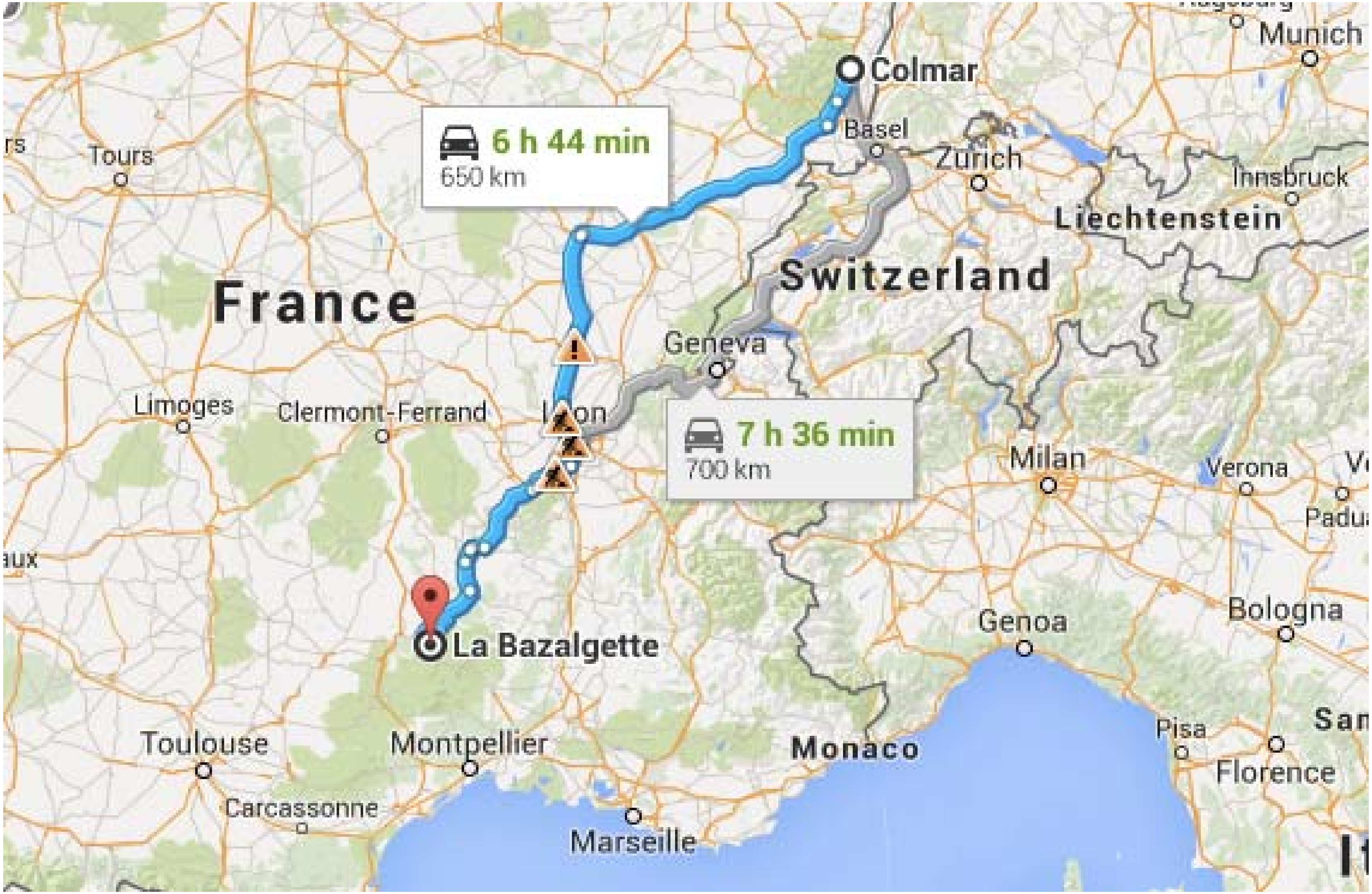

Colmar
6 h 44 min
650 km
Basel
Zurich
Munich
Innsbruck
Liechtenstein
Switzerland
Tours
France
Geneva
7 h 36 min
700 km
Limoges
Clermont-Ferrand
Milan
Verona
La Bazalgette
Genoa
Bologna
Toulouse
Montpellier
Monaco
Pisa
Florence
Carcassonne
Marseille

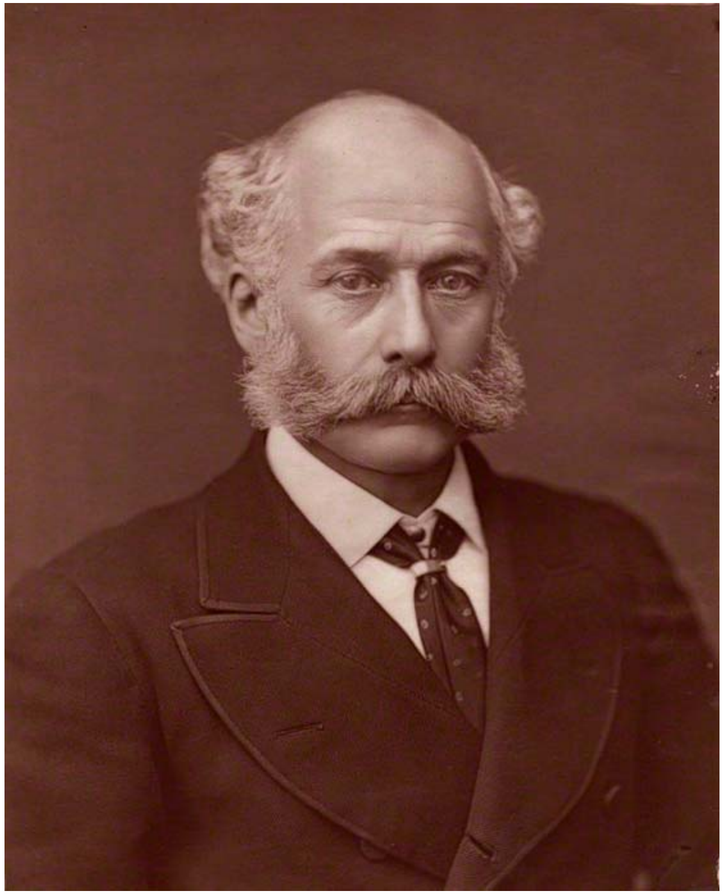
Table S2: Rare Oxbridge versus non-Oxbridge Surnames, 1800-29

Oxbridge		Non-Oxbridge	
Agassiz	Brickdale	Agnerv	Bodgett
Anquetil	Brooshooft	Allbert	Boolman
Atthill	Bunduck	Arfman	Bradsey
Baitson	Buttanshaw	Bainchley	Breckill
Barnardiston	Cantis	Bante	Callaly
Bazalgette	Casamajor	Barthorn	Capildi
Belfour	Chabot	Bavey	Carville
Beridge	Charretie	Bedborne	Cavet
Bleeck	Cheslyn	Bemond	Chanterfield
Boinville	Clarina	Berrton	Chesslow
Boscawen	Coham	Bideford	Chubham
Bramston	Conyngham	Bisace	Clemishaw

Source: 'Surnames and Social Mobility',Gregory Clark and Neil Cummins

http://www.econ.ucdavis.edu/faculty/gclark/ecn110a/readings/Surname%20Mobility%202013.pdf

Bazalgette



National Portrait Gallery

Surname: ‘BAZALGETTE’

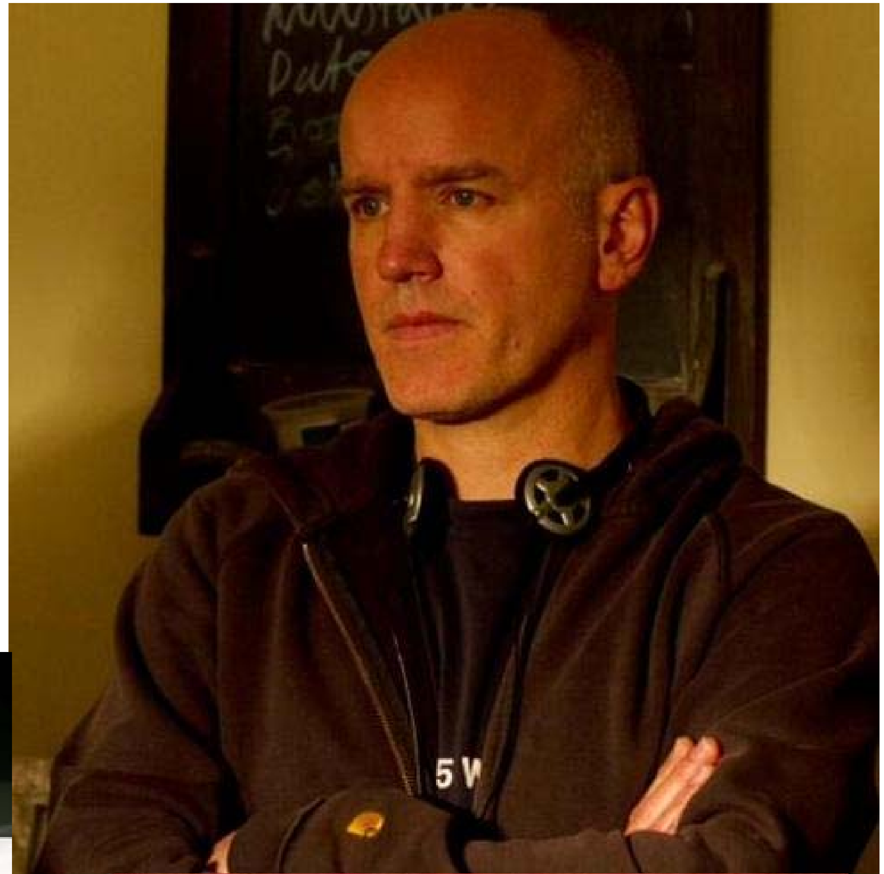
Forename	Surname	Address	Postcode
JANE	BAZALGETTE		BH1***
GUY	BAZALGETTE		BH1***
EMMA	BAZALGETTE		BH1***
ELIZABETH	BAZALGETTE		TW9***
VICTORIA	BAZALGETTE		LS1***
MARK	BAZALGETTE		DT1***
LUKE	BAZALGETTE		DT1***
MARIE	BAZALGETTE		DT1***
ELEANOR	BAZALGETTE		SW8***
LEE	BAZALGETTE		SA3***
ROBERT	BAZALGETTE		SA3***
ROBIN	BAZALGETTE		SA3***
RUTH	BAZALGETTE		PL1***
EMILY	BAZALGETTE		N15***
MARY	BAZALGETTE		BH1***
RICHARD	BAZALGETTE		BH1***

: ‘Surnames and Social Mobility’,Gregory Clark and Neil Cummins

http://www.econ.ucdavis.edu/faculty/gclark/ecn110a/readings/Surname%20Mobility%202013.pdf

‘using educational status in England 1170-2012 as an example, … the true underlying [intergenerational] correlation of social status is in the range 0.75-0.85. Social status is more strongly inherited even than height.’

This ‘stems from the nature of inheritance of characteristics within families. Strong forces of familial culture, social connections, and genetics must connect the generations. There really are quasi-physical “Laws of Inheritance.”’



You can't plan a career in too much detail': Simon Bazalgette (Mary Turner)

How are cities differentiated?

Lifestyle

Demographics

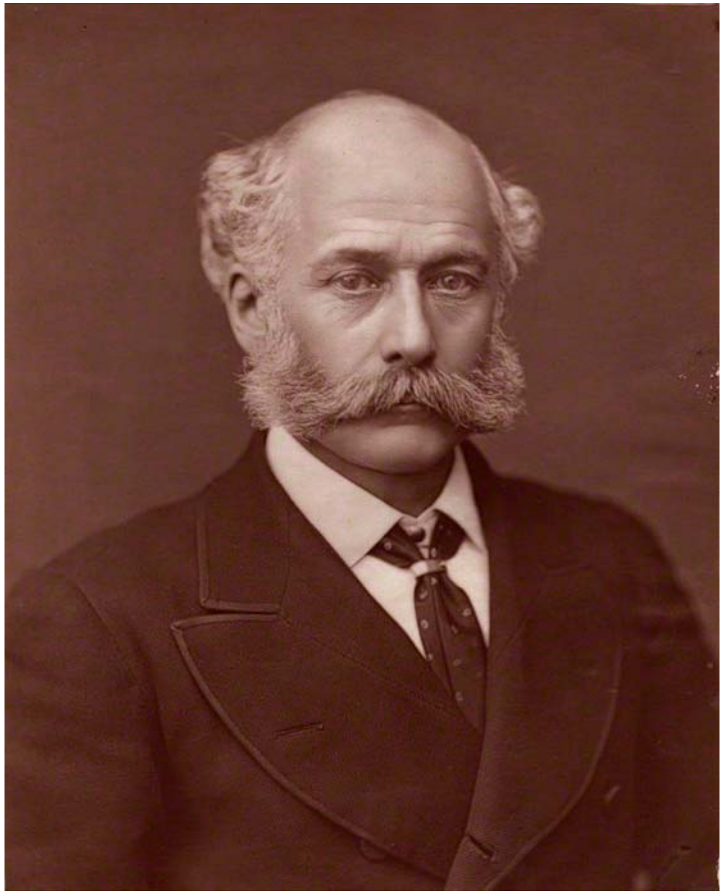
Housing & access

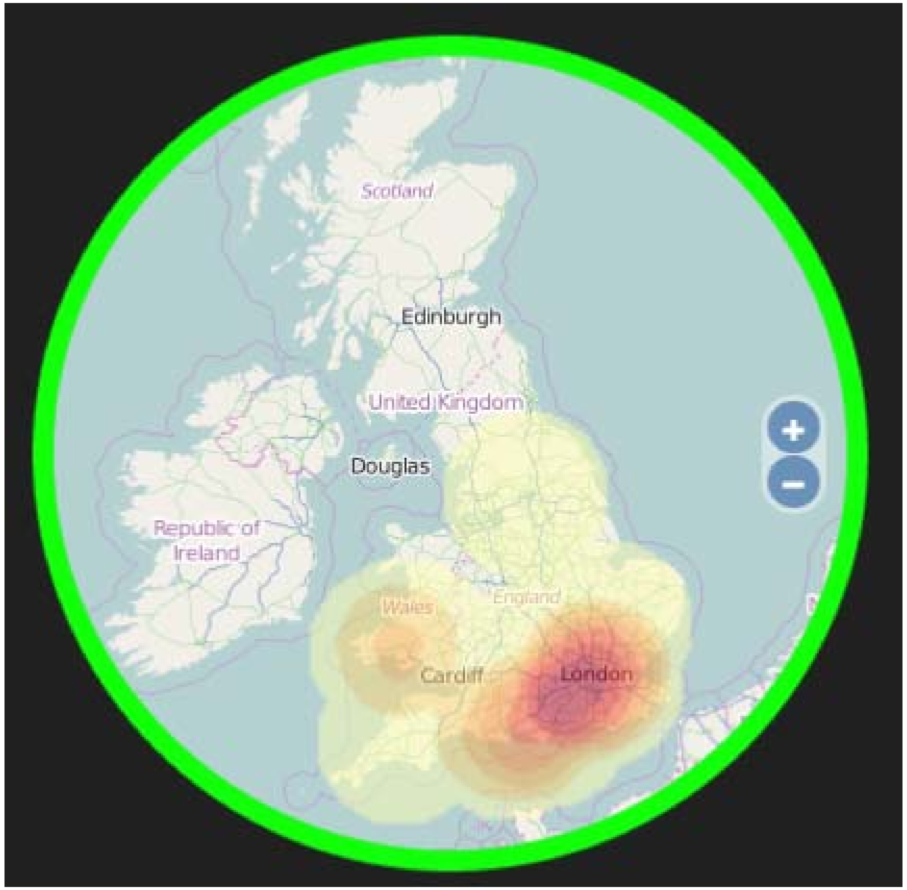
Socio-economics

Employment

Culture

Bazalgette

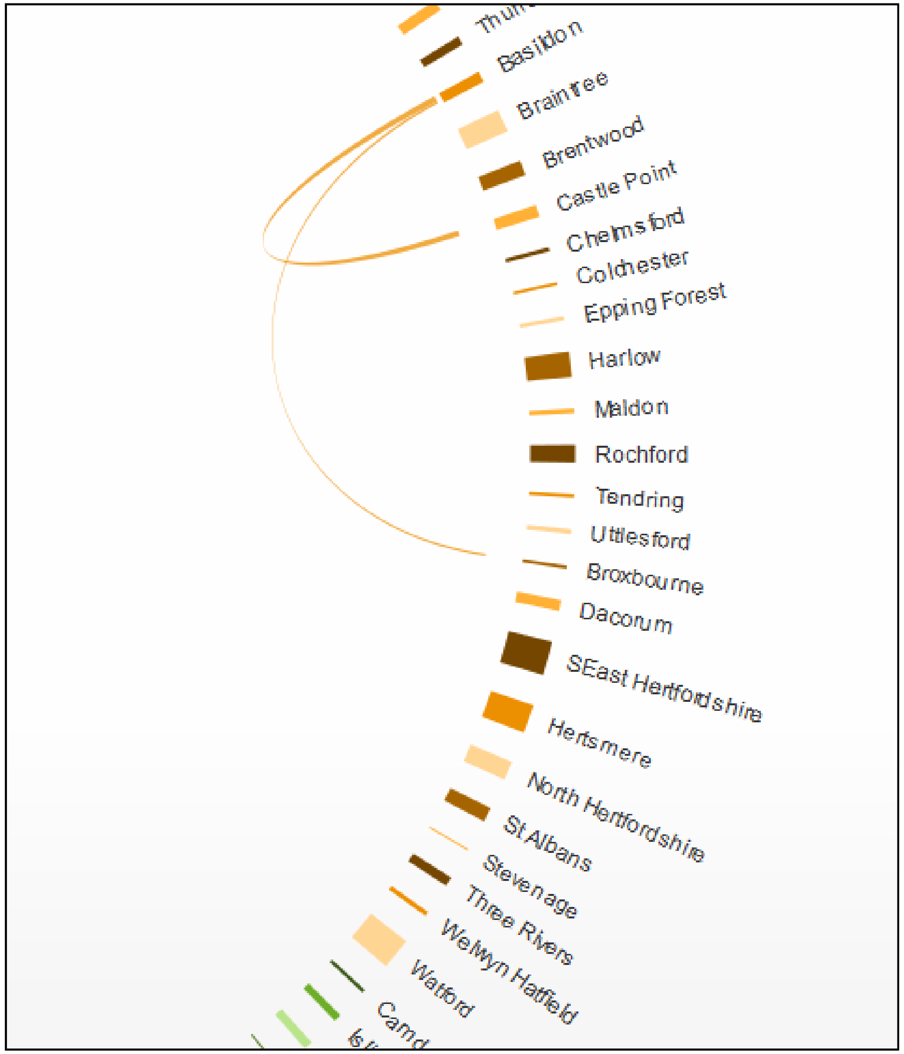




OSM, National Portrait Gallery

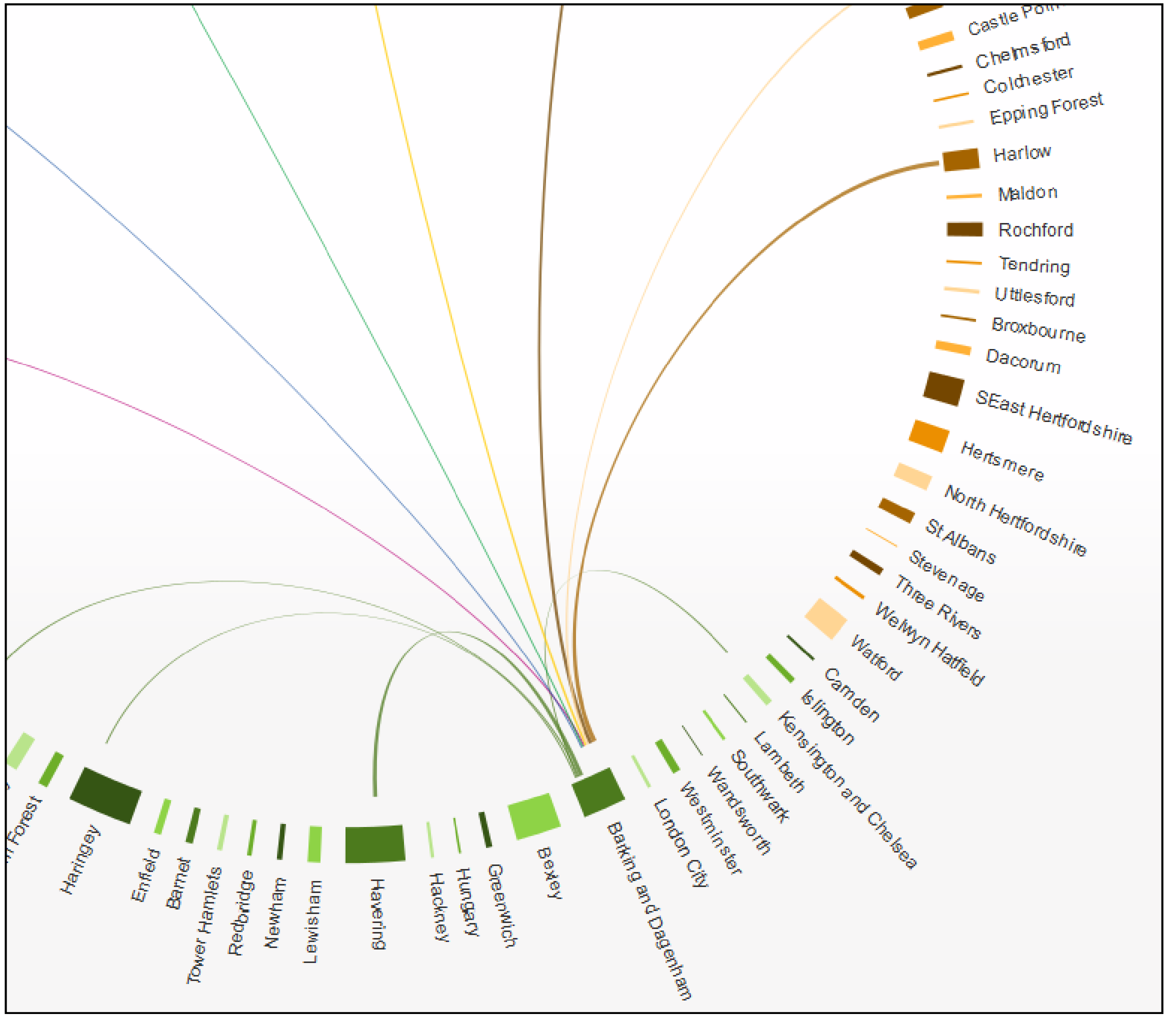
Initial Results

Movement out of Basildon:

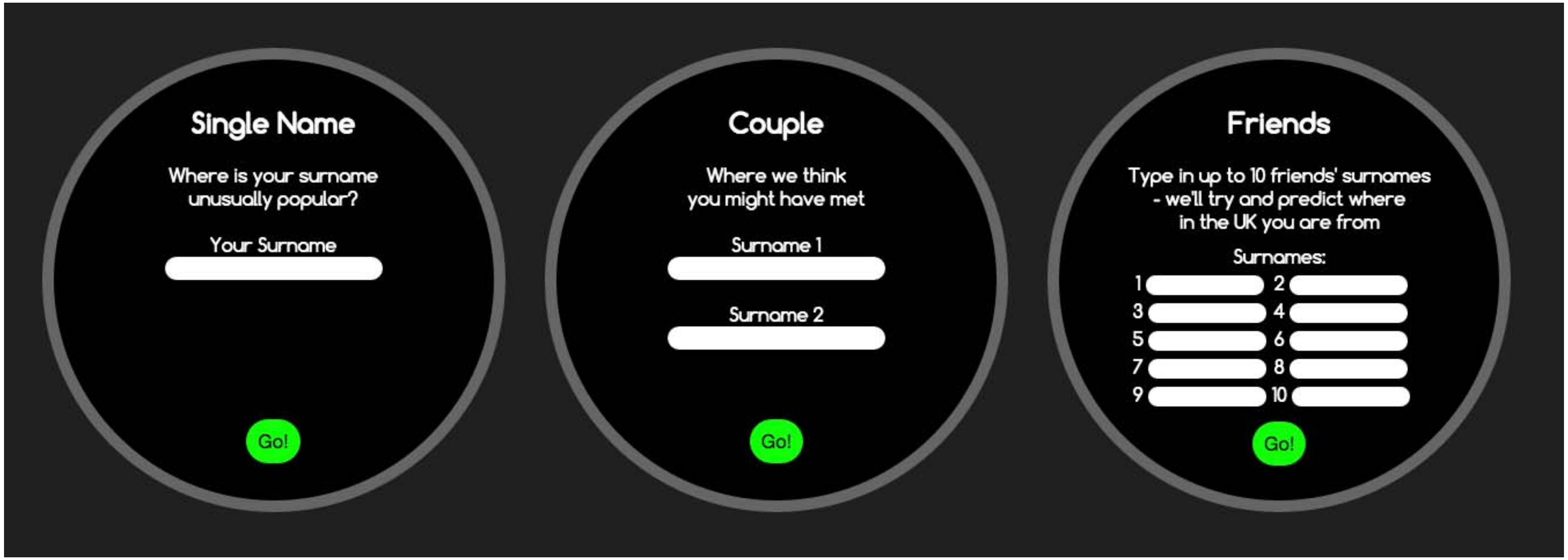


Initial Results

Movements into Barking and Dagenham:

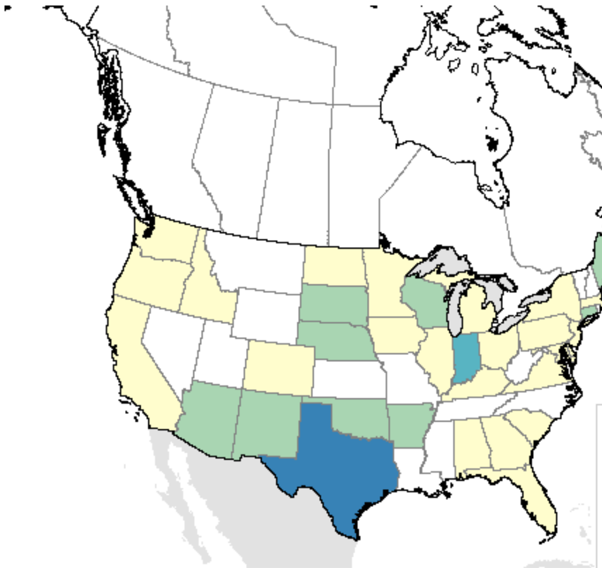


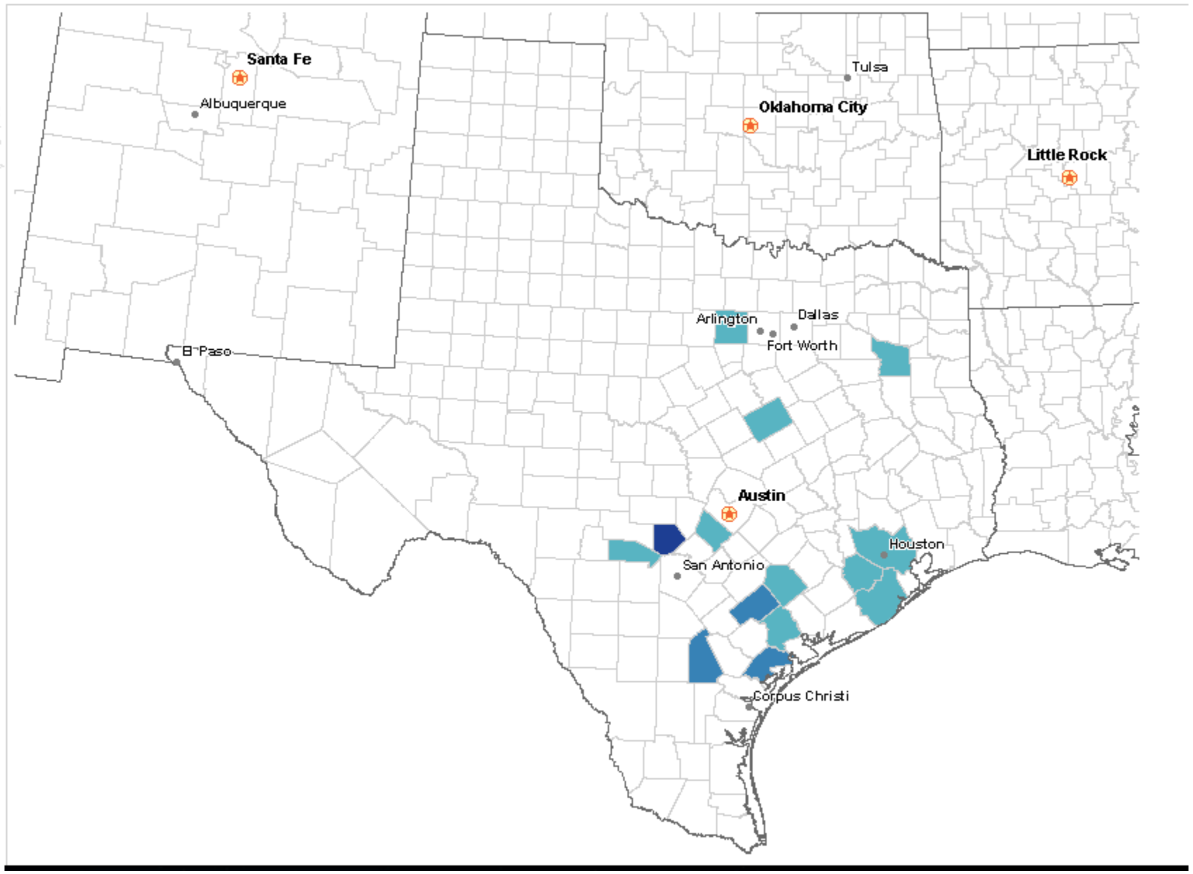
named


Single Name
Where is your surname
unusually popular?
Your Surname
Go!
Couple
Where we think
you might have met
Surname 1
Surname 2
Go!
Friends
Type in up to 10 friends' surnames
- we'll try and predict where
in the UK you are from
Surnames:
1
2
3
4
5
6
7
8
9
10
Go!

Conclusions

- Challenges of understanding geo-temporal data
  - Google flu trends; not the ‘End of Theory’
  - “N = all” ??; Google Translate in a stable unchanging world – but Twitter?
  - the “data exhaust” (Tim Harford); systematic bias
  - response rates and research methods: 2015 UK General Election
- Tesco ‘data mining’ (& Target false positives)
- Public acceptability of linkage based on anonymisation, not consent





Santa Fe
Albuquerque
Tulsa
Oklahoma City
Little Rock
Arlington
Dallas
Fort Worth
El Paso
Austin
Houston
San Antonio
Corpus Christi