Digital Data Infrastructures: interrogating the social media data pipeline.

Les Carr and Susan Halford, Web Science Institute, University of Southampton.

Based on a paper by Susan Halford, Mark Weal, Ramine Tinati, Les Carr and Catherine Pope in *New Media and Society,* http://journals.sagepub.com/doi/abs/10.1177/1461444817748953

Introduction: what do we mean by ethics?

- Minimalist (legal)

- Regulated practice: consent, confidentiality/anonymity & the right to withdraw

- Representational responsibility: methodological and epistemological

Social Media Data

- From social media practices and effects … to data
- An unexpected gift bringing rich research opportunities
- Enthusiasm - the telescope - the macroscope …

  ‘… it is as if the inner workings of private worlds have been pried open’ (Latour 2007)
- Scepticism:

  ‘[w]hatever value big data may have for “knowing capitalism”, its’ value to social science has … [f]or the present at least, to remain very much open to question’ (Goldthorpe 2016)

We have conducted a survey of 10,000 people.

We don't know how these 10,000 people were selected as the company that collected the data for us won't share this information, but we know that at some point they said something about Southampton or Portsmouth.

These may not be 10,000 unique individuals; we have no way of knowing. Some people may have filled the form in multiple times. Some of the forms (up to 20%) may have been filled in by computer algorithms. We don't know which is which.

Different people have filled in different variations of the form. We don’t know which is which.

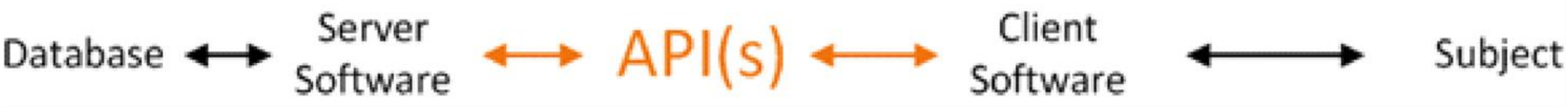
We have categorised people into those from Southampton and those from Portsmouth based on location information they have either supplied (very rarely), that we have inferred.

We have analysed the data using a small, not exhaustive, quite subjective set of positive and negative words to give each person a happiness rating from 1-10. Based on this we can say with a p=0.05 level of certainty that people living in Southampton are happier than people living in Portsmouth.

“For this paper, a number of interviews had been conducted, transcribed and printed by other researcher(s) for unknown reasons. The resulting printouts were torn into shreds, mixed up and dumped. I have used some of these found statements and interpreted them against my research questions.”

A middle path, between giving in and getting out (Gehl 2015)

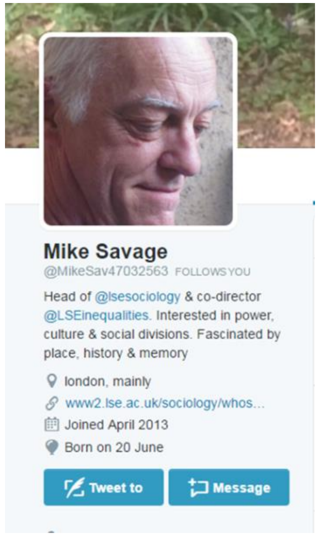
The Data Pipeline



- Sociotechnical
- Iterative
- Core to the generation of data
- Core to the circulation of data
- Methodological implications?

1: Population

- Demographics
- Location



Mike Savage

@MikeSav47032563 FOLLOWS YOU

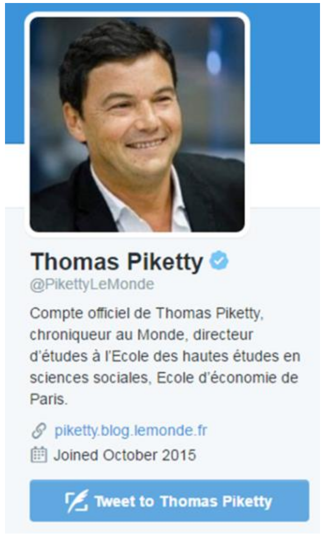
Head of @lsesociology & co-director @LSEinequalities. Interested in power, culture & social divisions. Fascinated by place, history & memory

london, mainly

www2.lse.ac.uk/sociology/whos...

Joined April 2013

Born on 20 June



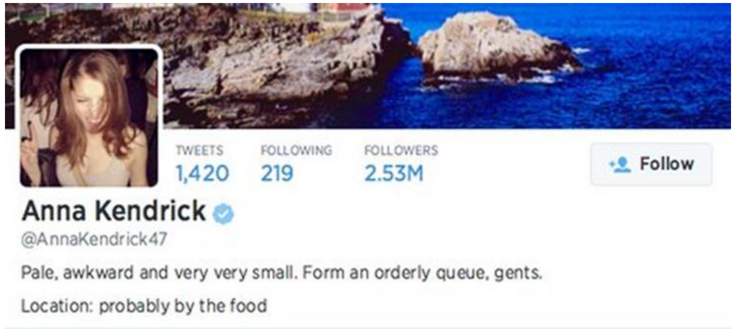
Thomas Piketty

@PikettyLeMonde

Compte officiel de Thomas Piketty, chroniqueur au Monde, directeur d'études à l'Ecole des hautes études en sciences sociales, Ecole d'économie de Paris.

piketty.blog.lemonde.fr

Joined October 2015



Anna Kendrick

@AnnaKendrick47

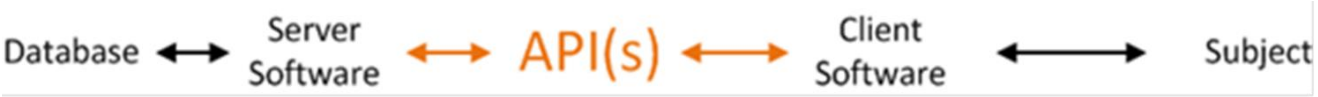
Pale, awkward and very very small. Form an orderly queue, gents.

Location: probably by the food

GPS location enabled - <3%

Jakarta 2.86%

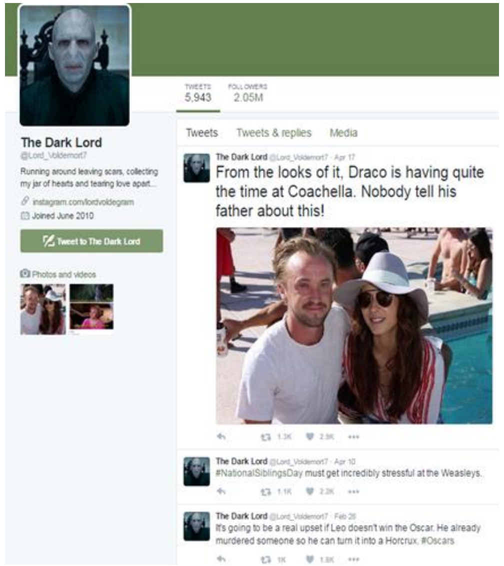
Moscow 0.77%



Database ↔ Server Software ↔ API(s) ↔ Client Software ↔ Subject

1: Population

- Demographics
- Location
- Users – sovereign individuals?



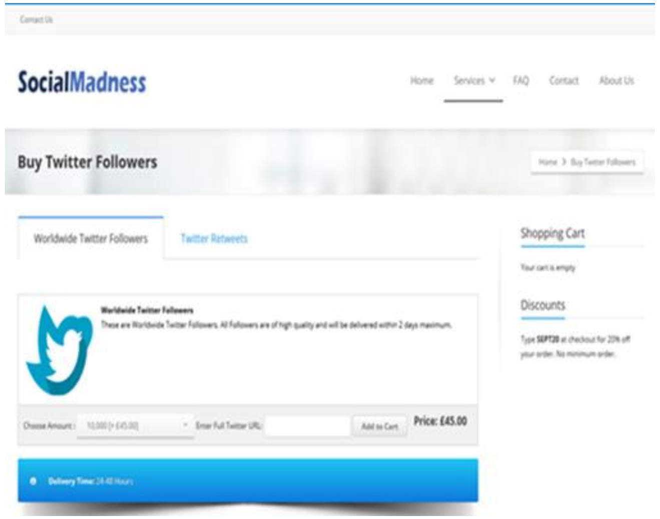
The Dark Lord
@Lord_Voldemort7
Running around leaving scars, collecting my jar of hearts and tearing love apart...
instagram.com/lordvoldegram
Joined June 2010

The Dark Lord @Lord_Voldemort7 · Apr 17
From the looks of it, Draco is having quite the time at Coachella. Nobody tell his father about this!

The Dark Lord @Lord_Voldemort7 · Apr 10
#NationalSiblingsDay must get incredibly stressful at the Weasleys.

The Dark Lord @Lord_Voldemort7 · Feb 28
It's going to be a real upset if Leo doesn't win the Oscar. He already murdered someone so he can turn it into a Horcrux. #Oscars

Add corporate account image



Buy Twitter Followers

Shopping Cart

Your cart is empty

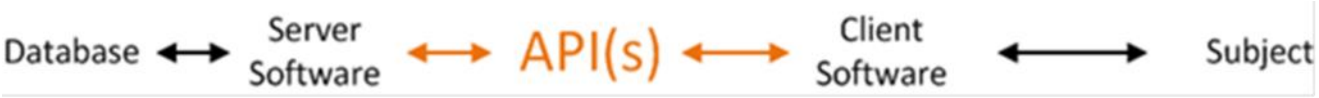
Discounts

Type SEPT20 at checkout for 20% off your order. No minimum order.

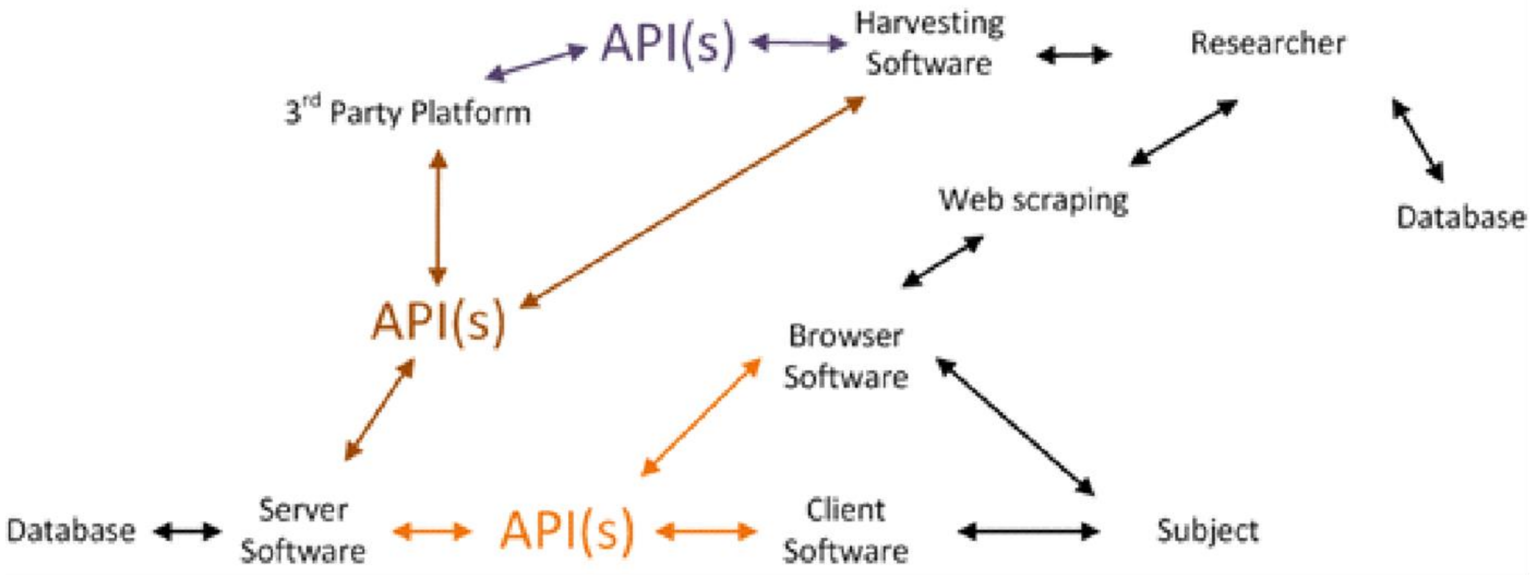
Worldwide Twitter Followers
These are Worldwide Twitter Followers. All Followers are of high quality and will be delivered within 2 days maximum.

Choose Amount: 10,000 (+ £45.00)
Enter Full Twitter URL:
Price: £45.00

Delivery Time: 24-48 Hours



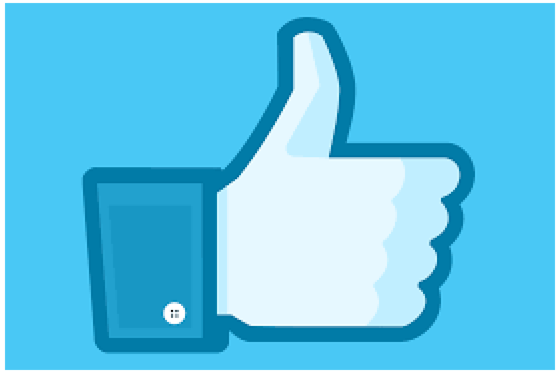
2: Sample



- How the data are harvested shapes the sample
- The API shapes the sample e.g. % data streams, real time/historic
- Rate limiting

3: Method

- Instruments for data collection
- Affordances
- For example: functionalities, data bases – shape data in specific ways over time



	Population	Sample	Method of data production
Database	Storage design and method shapes the types of information recorded about users.	Historic data storage decisions and technical query limitations may shape what data are included in samples.	Considerations of cost, performance and business requirements for data storage may shape what data are collected and stored and how.
Server Software	Determines who or what has access to the service, and what information is required to set up an account.	Server capacity may restrict data volume delivered; geographical location of server may affect data delivered.	Operates data management (e.g. spam removal and moderation, load balancing) shaping what data are collected.
API	APIs may not recognise all characters (languages) effectively; or be available to all operating systems/software development toolkits	A variety of differently structured samples may be available.	Defines the scope and volume of what data can be collected, stored and queried.
Harvesting Method	Harvesting methods construct different views of the populations. Web scraping may be more likely to access the population of currently active users, which could be different to the population accessed via historical searches using an API.	Web scraping will by-pass ‘official’ data samples, offering data from a sample of web pages. This sample may be affected by the ‘filter bubble’ of the person accessing the web pages. Use of third party data may introduce additional sampling effects.	Different harvesting methods have access to different types of data about the population and sample.
Client Software	Different clients may generate different information about the population. On some platforms you may know what client generated the content (this used to be the case on Twitter), on many though you can’t know this.	Some clients (apps) may receive more data than others (if harvesting through a client).	Different clients may produce distinctive forms of data and metadata e.g. some may add geographic data by default, some might link directly to shared or re-shared material.
Subject	Different subjects – human/non-human, demographically distinct – may characterise particular platform populations.	User activities may shape sampling methods (e.g. official samples may focus on central or highly active users.)	User practices and meanings shape the data generated and the claims that can be made from these.

The basic tenets of ethical practice include the fundamental rights of human dignity, autonomy, protection, safety, maximization of benefits and minimization of harms, or, in the most recent accepted phrasing, respect for persons, justice, and beneficence.

Conclusion

- Recognise the limits of what we can and can’t know about social media data

- Key steps

(1) Transparency
(2) Consider implications of data construction for research questions
(3) Knowledge claims
(4) Creative data assemblages