Big data: the coming crisis of governance and practice?

Linnet Taylor

University of Amsterdam

14 March 2016

Geo-information: proprietary and ubiquitous

- Digitisation and ‘datafication’ provide new data and new analytical possibilities – 7.4bn mobile connections, 5.5bn in LMICs, 1bn in Africa

- Smartphones change access to geo-information
  - GPS, cell tower signals, wifi signals, Bluetooth sensors, IP addresses, network environment data

- ‘LMICs will provide the majority of geolocated digital data by 2020’ (Manyika et al., 2011)

‘The god’s eye view’ (Pentland 2011)

- Monitoring and surveillance
  - Human rights
  - Epidemiology
  - Nowcasting
  - Crowdsourcing

- Sorting and categorising
  - Biometrics & welfare delivery
  - Poverty mapping

- Identifying trends
  - Mobility, urbanisation
  - Refugee flows
  - Financial flows & agricultural learning

How is proprietary data shared?

- Data analysts
  - – Basic research (academia)
  - – Commercial research (private sector)

- Data intermediaries
  - – For-profit – Mezuro
  - – Not-for-profit – Flowminder, DataPop, UN Global Pulse
  - – Mixed – Grameen applab

Ethics and geo-data
1) can we understand the data?

A story about ground truth

D4D winning paper, ‘development’ category:

Berlingerio et al. (2013) *AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data*

- Problem: how to address inefficiencies in Abidjan’s transport system

‘If transit agencies could have an effective tool to **quantify the travel demand**, as well as recommendations on **how to best design** the transit network, cities would be able to better support travelers’ mobility demand through a **regulated and efficient public transport system**.’

(Berlingerio et al. 2013)

Abidjan: the god's eye view

Abidjan mobile phone traffic through Orange's network

Existing transport use (thicker lines = busier)

Optimised transport network (red line: new route)

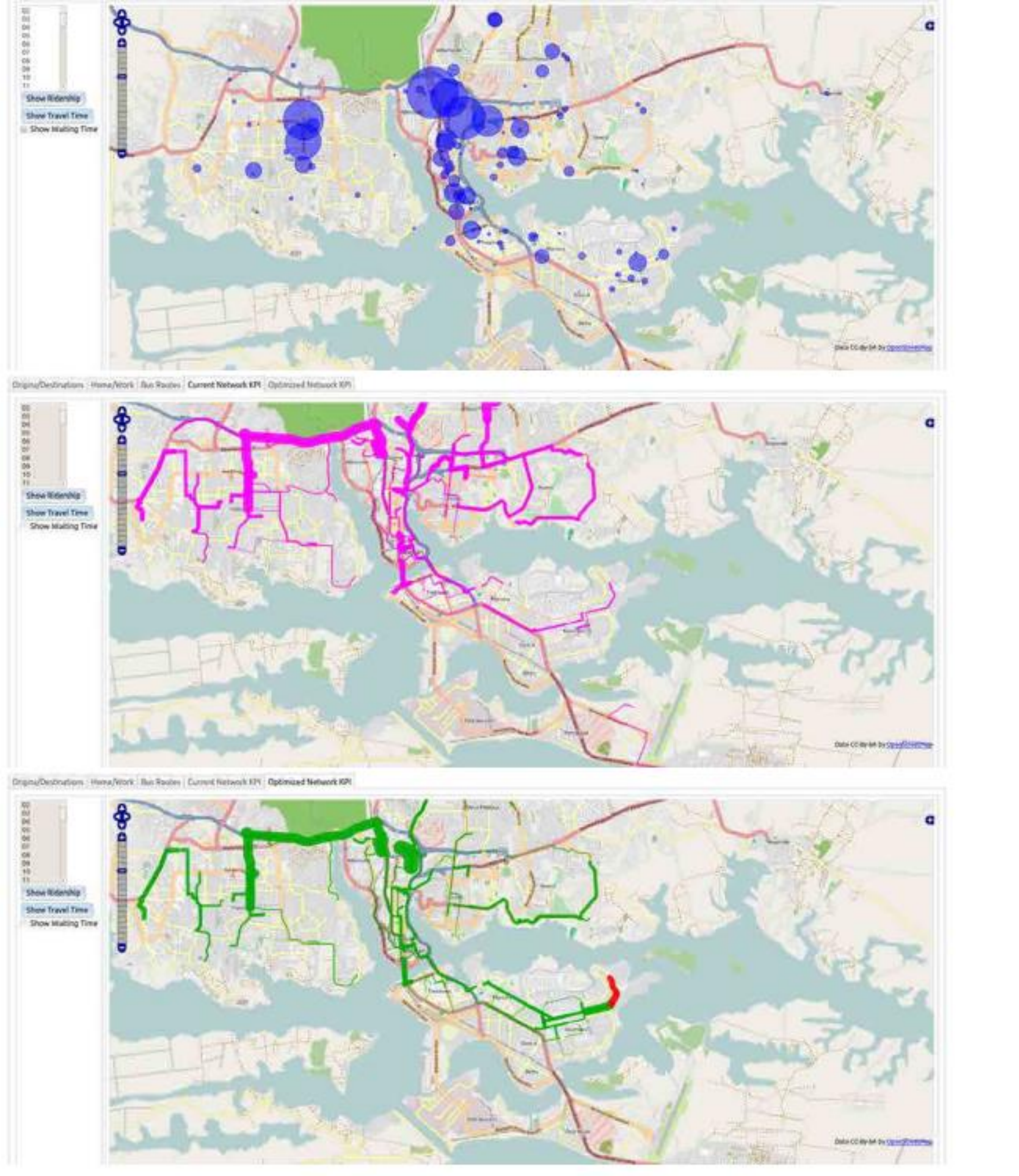


Figure 16: Current SOTRA network. Top: radius of circles is proportional to waiting times at stops. Middle: line width is proportional to the ridership of the line. Bottom: SOTRA network and additional routes (line width is proportional to the ridership of the line).

Bias in the sample

- Phone data represents commuters on all forms of transport; maps relate only to SOTRA (Société des transports Abidjanais), i.e. official buses.
- '…accurate SOTRA transit route and schedule information was not available. We then decided to leverage all available Web information to extract **reasonable bus stop location** as well as route shape information. Unfortunately we were **not able to fully validate** the extracted transit network information. We hope this could be achieved in the near future with the help of the local authorities, and potentially with citizen engagement.'

(Berlingerio et al. 2013)

Abidjan’s public transport system

- SOTRA: 3 services, 10-30% of Abidjan’s transport (Lombard 2006), further deregulated over last decade.
- Gbakas: private minibuses
- Taxis collectifs
- ‘Véhicules banalisés’ (informal taxis)
- Ndiaga Ndiaye: intercity buses

Abidjan transport: the local view

SOTRA buses

Travel problems due to civil war?

© picture-alliance/dpa

Official taxis

Gbakas

Reframe the data, reframe the problem

- Data scientists’ assumption of vertical, monopolistic governance:
  - ‘an existing transit network’; ‘a fleet’; ‘a public transport operator’
- vs. local reality
  - layered governance models: state, city, firms, unions, informal assns
  - no unified transport system or authority (SOTRA 0.07% city-owned)
  - weak vertical but strong horizontal governance after liberalisation
  - Recent civil war created travel bottlenecks
- Was the problem inefficiency or informality?
- Alternate interpretation: flexible, responsive transport system with inefficiencies due to infrastructure, no centralised transport planning.
- Problem of truthiness (Cherlet 2014)

Garbage In, Garbage Out?

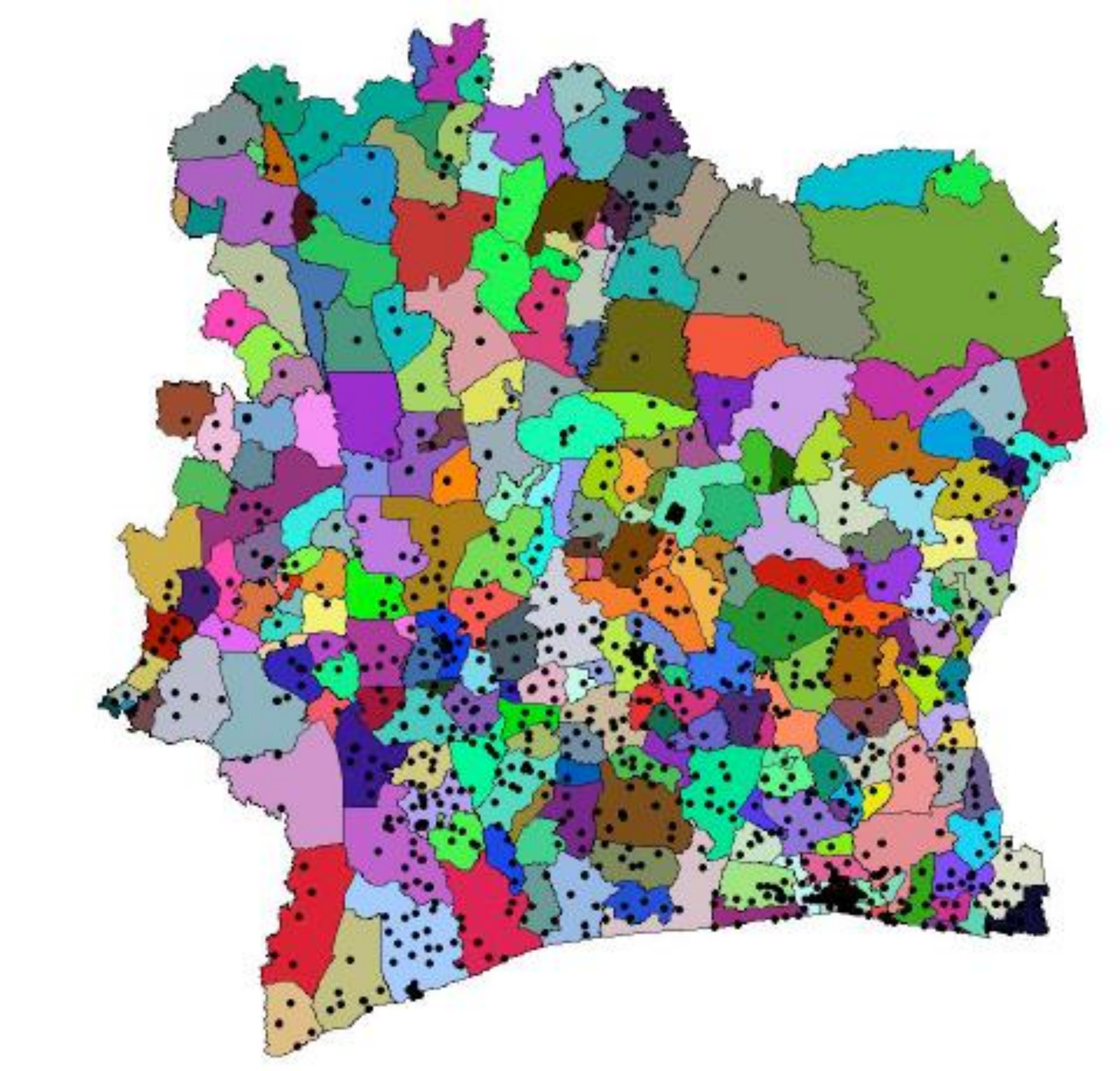
‘Big data are not about society, they are about users and markets.’ (Shearmur 2015)

Biases: socioeconomic, gender, ethicity, connectivity, smartphones (GIGO)

Big data fosters epistemological determinism (truthiness)

Orange’s cell towers in Côte d’Ivoire, 2013

(sub-prefecture administrative regions)



Distribution of cell phone antennae is denser in urban areas than rural locations, making it easier to detect activity in urban/denser populated areas.

Source: Blondel et al. 2013

Population estimation error

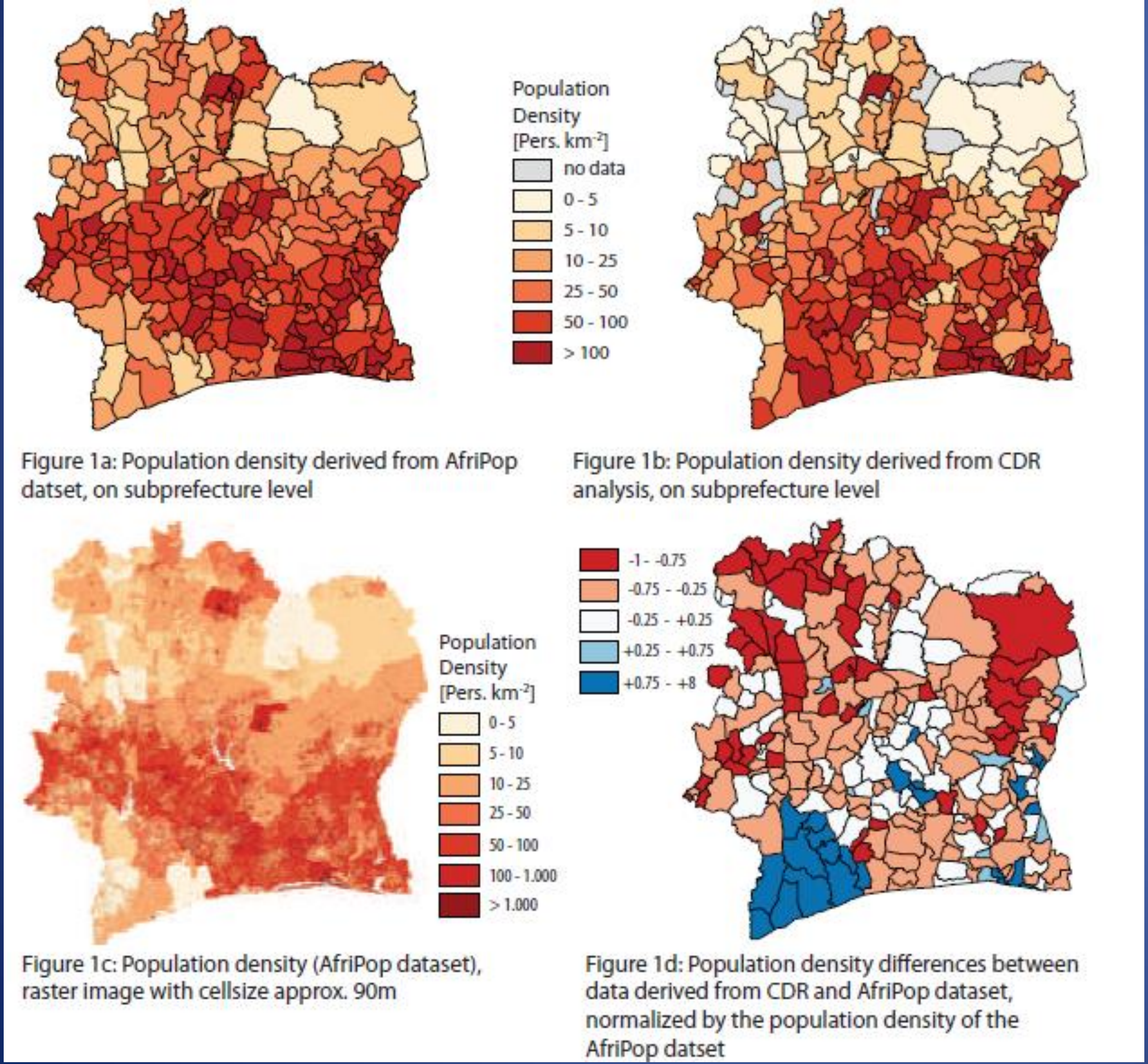


Figure 1a: Population density derived from AfriPop datset, on subprefecture level

Figure 1b: Population density derived from CDR analysis, on subprefecture level

Figure 1c: Population density (AfriPop dataset), raster image with cellsize approx. 90m

Figure 1d: Population density differences between data derived from CDR and AfriPop dataset, normalized by the population density of the AfriPop datset

Over/underestimation of population according to cellphone activity (figure 1d) follows unequal distribution of cell towers (see previous slide)

(Sterly et al. 2013)

Ethics and geo-data: 2) who is accountable?

- The problem of policy-relevant research with sensor data

‘There is **intelligence-grade situational awareness** in the case of sensors: what can be used to document a human rights abuse can also be used to target an artillery strike.’ (Nathaniel Raymond, HHI, 25.2.2015)

3) What are the guidelines?

- Fair Information Practice Principles:
  - Notice/awareness
  - Consent
  - Access/participation
  - Integrity/security
  - Enforcement/redress

- UN Data Revolution rights:
  - Right to be counted
  - Right to an identity
  - Right to privacy and to ownership of personal data
  - Right to due process
  - Freedom of expression
  - Right to participation
  - Right to non-discrimination and equality
  - Principles of consent (UN 2014)

NB: None of these work with big data

4) Which institutions can help?

- Big data 1: IRBs 0
  - FB experiment, "**Experimental** evidence of massive-scale emotional contagion through social networks" (2014)
  - Flowminder – IRBs unwilling to act across borders and tech regimes
  - Ebola in Liberia: the Liberian MoH's Internal Review Board shut down during the crisis so that researchers had to rely on their own IRBs
- Orange D4D ethics committees are an example of successful self-regulation by a corporation

4) Can’t we just anonymise it?

- Anonymisation vs. de-identification
  - ‘there are no perfect ways to de-identify data and there probably never will be.’ (Kendall et al. 2014)
  - Four spatio-temporal points are enough to uniquely identify 95% of individuals (de Montjoye et al. 2013)
  - Uniqueness of mobility traces decays approximately as the 1/10 power of their resolution (de Montjoye et al. 2013)

‘Even if you are looking at purely anonymized data on the use of mobile phones, carriers could predict your age to within in some cases plus or minus one year with over 70 percent accuracy. They can predict your gender with between 70 and 80 percent accuracy. One carrier in Indonesia told us they can tell what you’re religion is by how you use your phone. You can see the population moving around.’

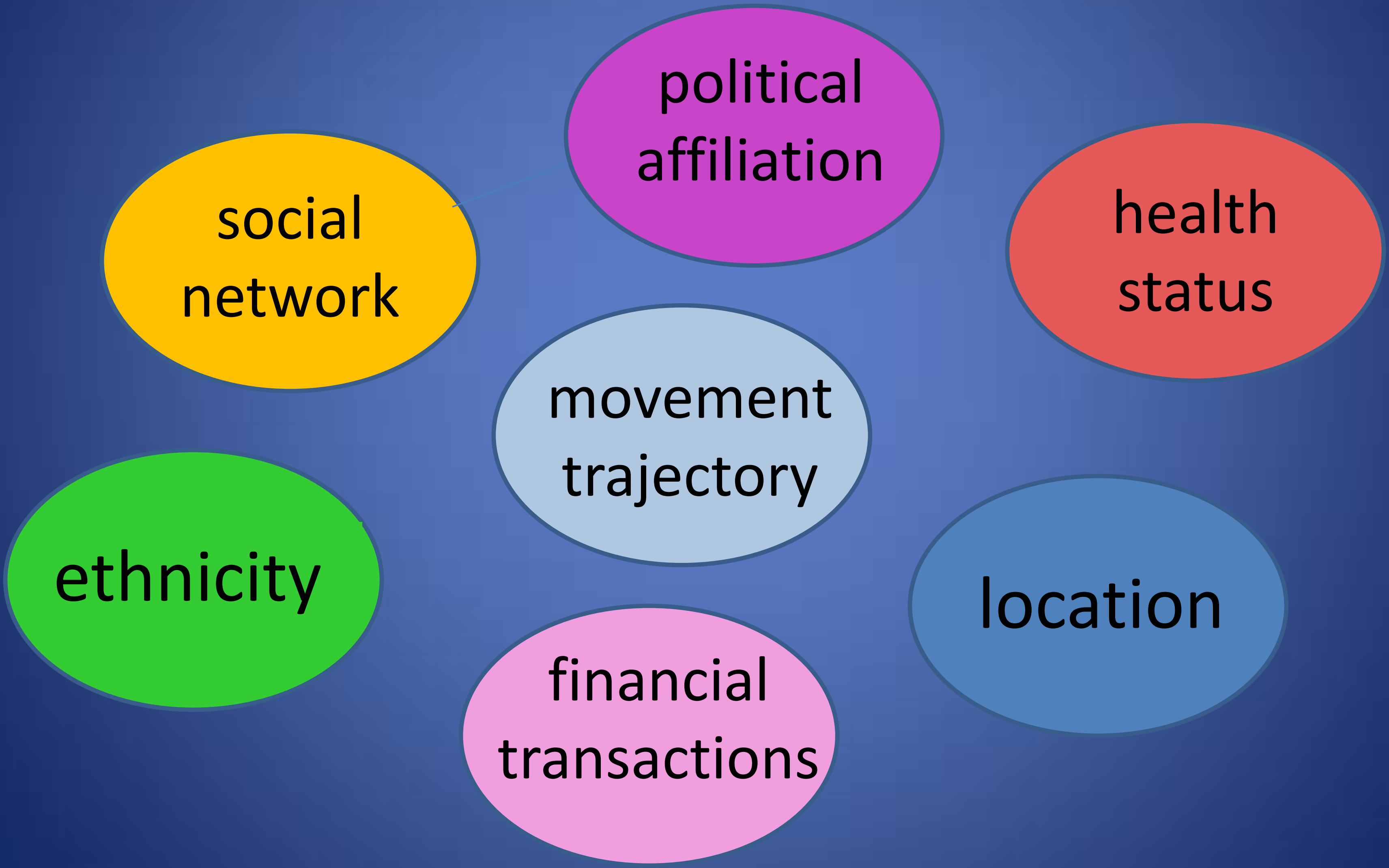
-- Robert Kirkpatrick UN Global Pulse, 2012

So? What next?

Addressing the grey areas: PII vs constitutive information

- Ethical big data research requires a distinction between ‘dead personal information’ vs ‘ontologically constitutive information’ (Floridi 2013)
- Dead personal info: name, rank, serial number
- Ontologically constitutive information: a person’s Google search history over the last month; where their car was parked last night

Ontologically constitutive information



political affiliation

social network

health status

movement trajectory

ethnicity

location

financial transactions

Personally identifiable vs. Demographically identifiable information

- Big data can proxy for categories where data may be restricted (e.g. political affiliation, ethnicity, religion, home/work location)

- Calling patterns or an active leader make groups trackable across space as networks (Sharad and Danezis 2013)

‘the volunteered information of the few can unlock the same information about the many’ (Barocas and Nissenbaum 2014)

- Nesting a survey within a large, anonymised mobile dataset allows researchers to de-anonymise the larger dataset in terms of group characteristics, i.e. age, gender, profession, employment status… (Blumenstock 2012) ***N.B. Rwanda study***

Big data, group privacy?

- Group harms
  - – Ethnic/religious/economic/political persecution
  - – Tracking/restriction of free movement
  - – ‘Aiding surveillance’ (Privacy International 2014)
  - – Identifying activist networks (McKinnon 2012)
  - – Manipulation (e.g. elections, moods)

Questions toward an ethical framework

- Consent, choice and accountability:
  - What does user consent mean under conditions of limited technological awareness and no accountability?
  - How can the FIPPs be reworked for big data?
- Individuals **plus** groups:
  - How do place/population characteristics affect the definition of data misuse?
  - How to move from compliance to doing no harm?
- Beyond compliance

References

- Barocas, S., & Nissenbaum, H. (2014). *Big data's end run around anonymity and consent* (pp. 44-75). Cambridge University Press, NY
- Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., & Sbodio, M. L. (2013). 'AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data'. In *Machine Learning and Knowledge Discovery in Databases* (pp. 663-666). Springer Berlin Heidelberg.
- Cherlet, J. (2014). Epistemic and technological determinism in development aid. *Science, Technology & Human Values*
- de Montjoye Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*.
- Floridi, L. (2013) *The Ethics of Information*. Oxford: OUP
- Kendall, Jake; Kerry, Cameron F.; and Montjoye, Alexandre de. "Enabling Humanitarian use of Mobile Phone Data," Issues in
- Technology Innovation (online) November 2014: http://www.brookings.edu/~/media/research/files/papers/2014/11/12-enablinghumanitarian- mobile-phone-data/brookingstechmobilephonedataweb.pdf.
- Lombard, J. (2006). 'Enjeux privés dans le transport public d'Abidjan et de Dakar'. *Géocarrefour*, 81(2), 167-174.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A. (2011). 'Big data: the next frontier for innovation, competition and productivity'. Washington DC: McKinsey Global Institute.
- Pentland, A. (2011). Society's nervous system: building effective government, energy, and public health systems. Pervasive and Mobile Computing 7(6): 643-659.
- Sterly,H.; Hennig, B; Dongo, K. (2013) "Calling Abidjan" – Improving Population Estimations with Mobile Communication Data (IPEMCODA). Retrieved from ResearchGate.net
- United Nations. (2014). *A world that counts*. New York.