Big data for beginners

An introduction for developmental professionals and researchers

Sriganesh Lokanathan, LIRNE*asia*

Research relevant to broadband policy and regulatory processes

New Delhi, India

21 August 2014

LIRNEasia
Pro-poor. Pro-market.

FORDFOUNDATION

IDRC CRDI
Canada

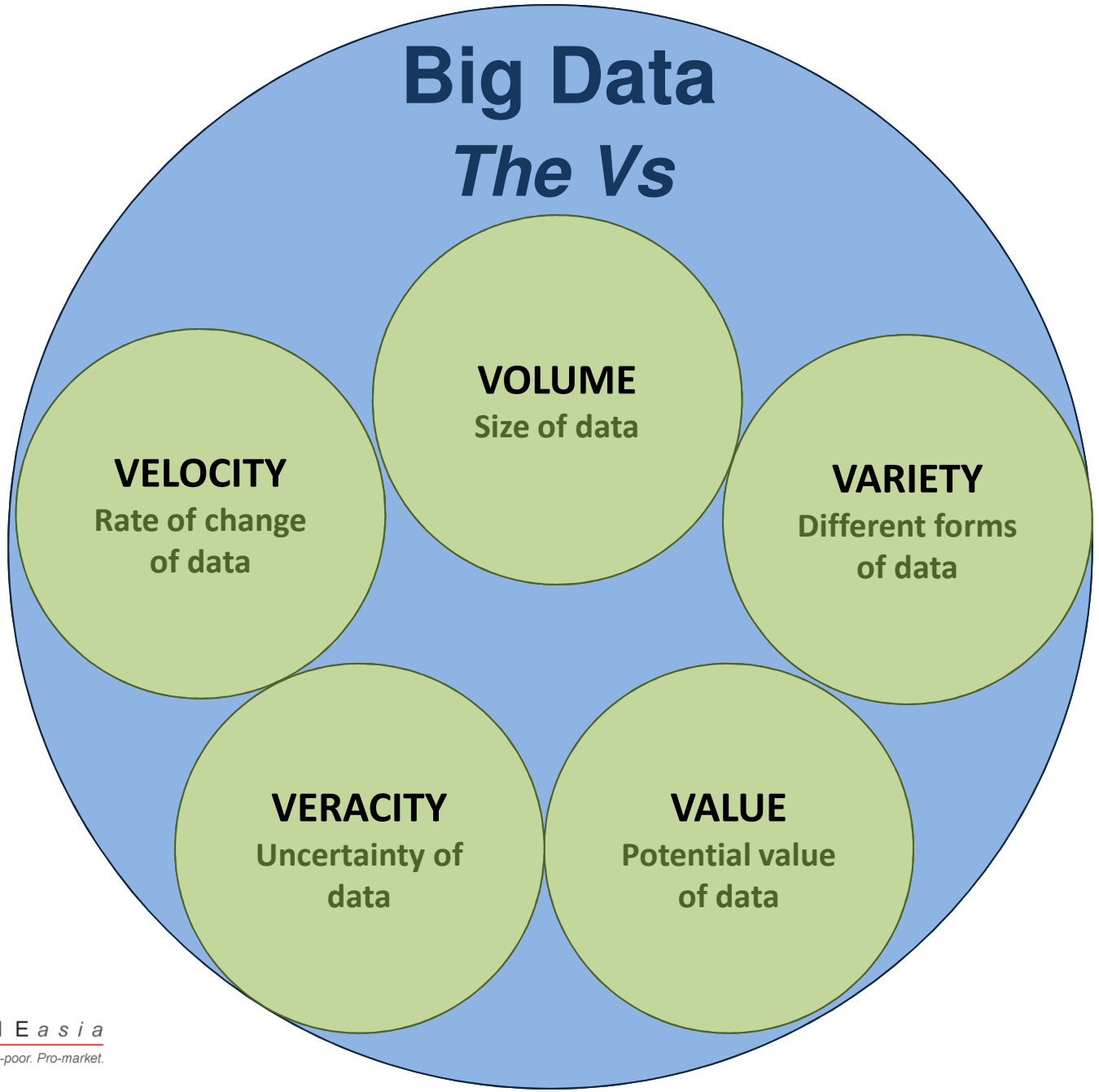
This work was carried out with the aid of a grant from the International Development Research Centre, Canada and the Department for International Development UK..

UKaid
from the British people

Learning outcomes for today

- Some understanding of what big data means
- An appreciation for some of the developmental opportunities from leveraging big data using different sources
- A brief overview of the spillover effects of the big data phenomenon
- An appreciation of some of the analytical and other challenges concerning the use of big data for development

What is big data?



LIRNEasia
Pro-poor. Pro-market.

What has facilitated the rise of big data?

- Vast drops in the cost of storing and retrieving information
- Exponential growth in computer power and memory
  - data can reside in persistent memory instead of disk and tape
- Major improvements in techniques for performing machine learning and reasoning

Sources of big data

- Administrative data
  - E.g. digitized medical records, insurance records, tax records, etc.
- Commercial transactions
  - E.g. Bank transactions, credit card purchases, supermarket purchases, online purchases, etc.
- Sensors and tracking devices
  - E.g. road and traffic sensors, climate sensors, equipment & infrastructure sensors, mobile phones, satellite/ GPS devices, etc.
- Online activities/ social media
  - E.g. online search activity, online page views, blogs/ FB/ twitter posts, online audio/ video/ images, etc.

Big data isn't necessarily a new area: the census is the original big data

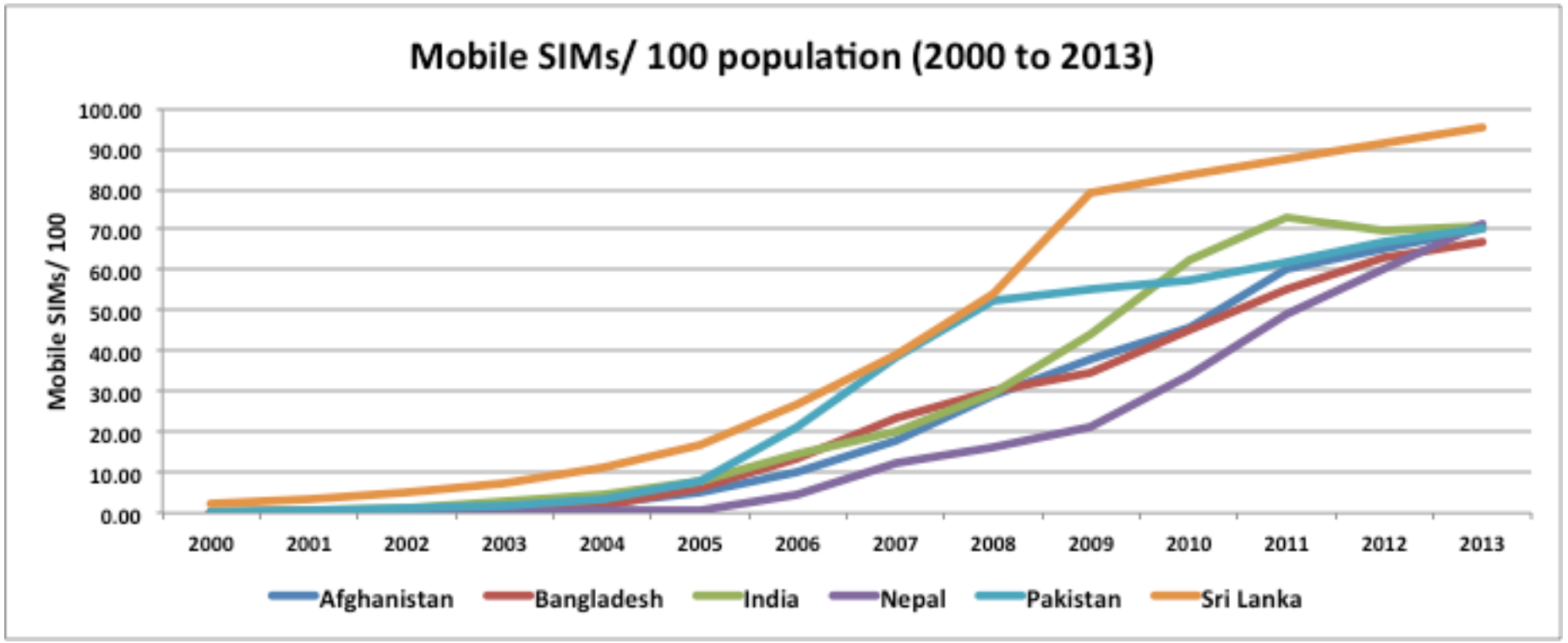
- First recorded census occurred in Egypt (circa 3340 BC)
  - But the tools didn't exist to fully exploit them
- The first tools invented for the 1890 US census
  - It took 7 years to process the 1880 census data
  - Herman Hollerith invents the tabulating machine to process the 1890 census data

Source: Adam Schuster

If we want comprehensive coverage of the population, what are the sources of big data in developing economies?

- Administrative data?
  - E.g. digitized medical records, insurance records, tax records, etc.
- Commercial transactions?
  - E.g. Bank transactions, credit card purchases, supermarket purchases, online purchases, etc.
- Sensors and tracking devices?
  - E.g. road and traffic sensors, climate sensors, equipment & infrastructure sensors, mobile phones, satellite/ GPS devices, etc.
- Online activities/ social media?
  - E.g. online search activity, online page views, blogs/ FB/ twitter posts, online audio/ video/ images, etc.

Currently only mobile network big data has the widest possible population coverage



LIRNEasia
Pro-poor. Pro-market.

LIRNE*asia*'s Big Data for Development (BD4D) Research

- LIRNE*asia* has negotiated access to **historical and anonymized** telecom network big data from multiple operators in Sri Lanka
- In the current research cycle we are
  - conducting exploratory research on answering a few social science questions related to mobility and connectedness
  - developing a framework with privacy and self-regulatory guidelines for the collection, use and sharing of mobile phone data.

- http://lirneasia.net/projects/bd4d/

The data sets

- Multiple mobile operators in Sri Lanka have provided LIRNE*asia* access to 4 different types of meta-data:
  - Call Detail Records (CDRs)
    - Records of calls, SMS-es, Internet access
  - Airtime recharge records
- Data sets do not include any Personally Identifiable Information (PII).
  - All phone numbers are anonymized and
  - LIRNE*asia* does not maintain any mappings of identifiers to original phone numbers

What are some types of big data captured by mobile network operators?

- Call Detail Record (CDR)
  - Records of all calls made and received by a person created mainly for the purposes of billing
  - Similar records exist for all SMS-es sent and received as well as for all Internet sessions

Calling Party Number	Called Party Number	Caller Cell ID	Call Time	Call Duration
A24BC1571X	B321SG141X	3134	13-04-2013 17:42:14	00:03:35

  - The Cell ID in turn has a lat-long position associated with it.

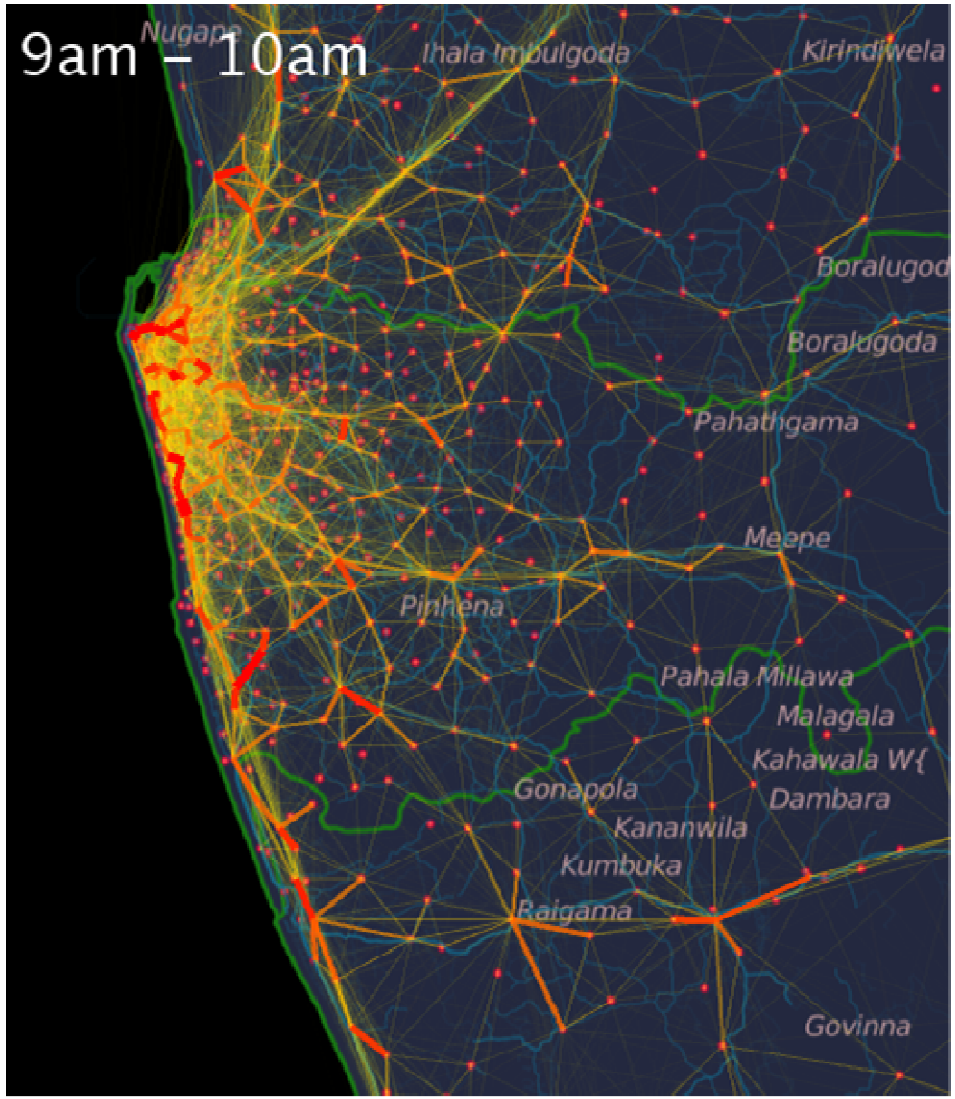
What are some types of big data captured by mobile network operators (contd.)?

- **Airtime reload records**
  - Records of all airtime reloads performed by prepaid SIMs
  - Each row corresponds to a record of one person's activity:

Number	Type of recharge	Starting balance	Amount	Call Time
A24BC1571X	CARD	0.41	50	13-04-2013 17:42:14

How can we use mobile network big data for development?

We can understand the mobility patterns of the population



Low → High

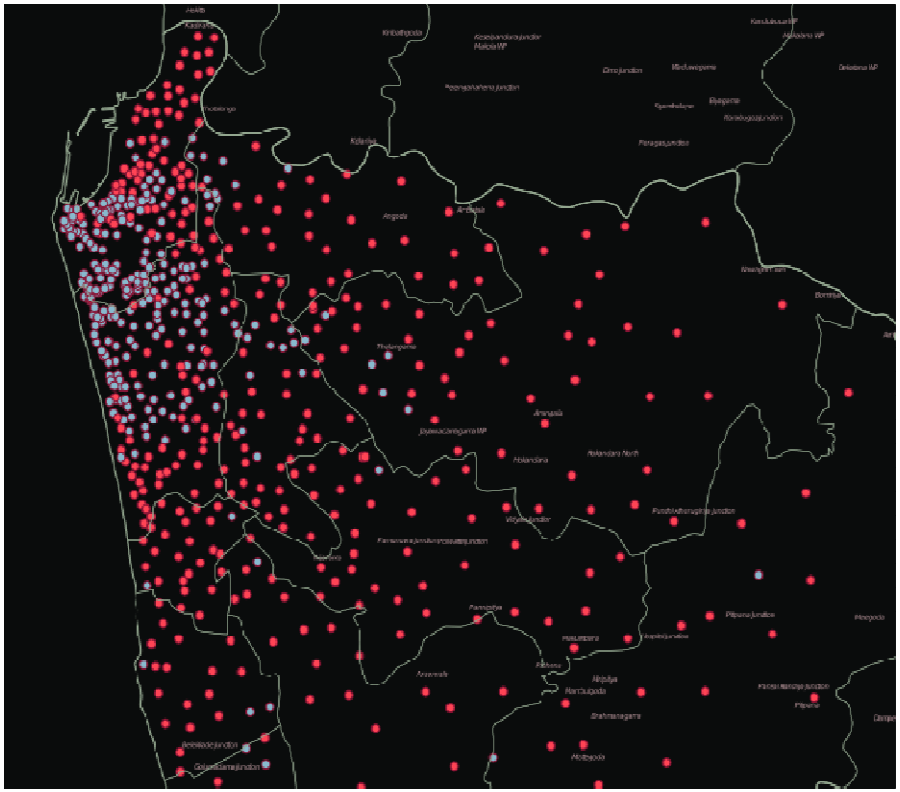
Volume of people

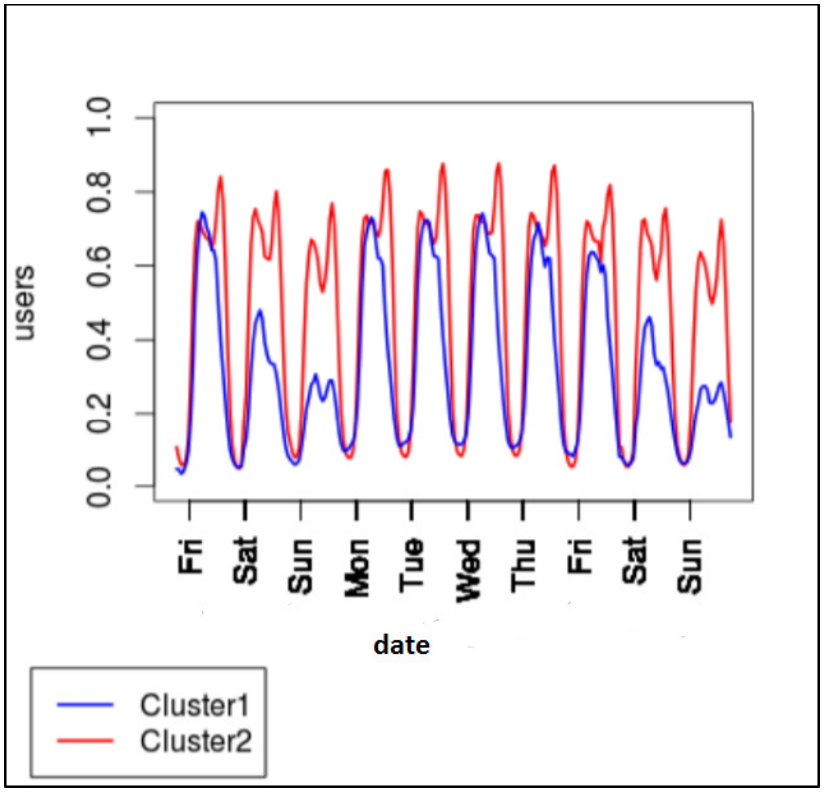
LIRNEasia
Pro-poor. Pro-market.

We can understand land use characteristics

- People leave digital traces when they use communication devices.
- Diurnal patterns of user activity in different locations can be used to classify land use characteristics
  - Euclidean distance between two time series is used to cluster base stations in an unsupervised manner using k-means algorithm

With even a simplistic clustering into just two regions we can clearly see the commercial areas



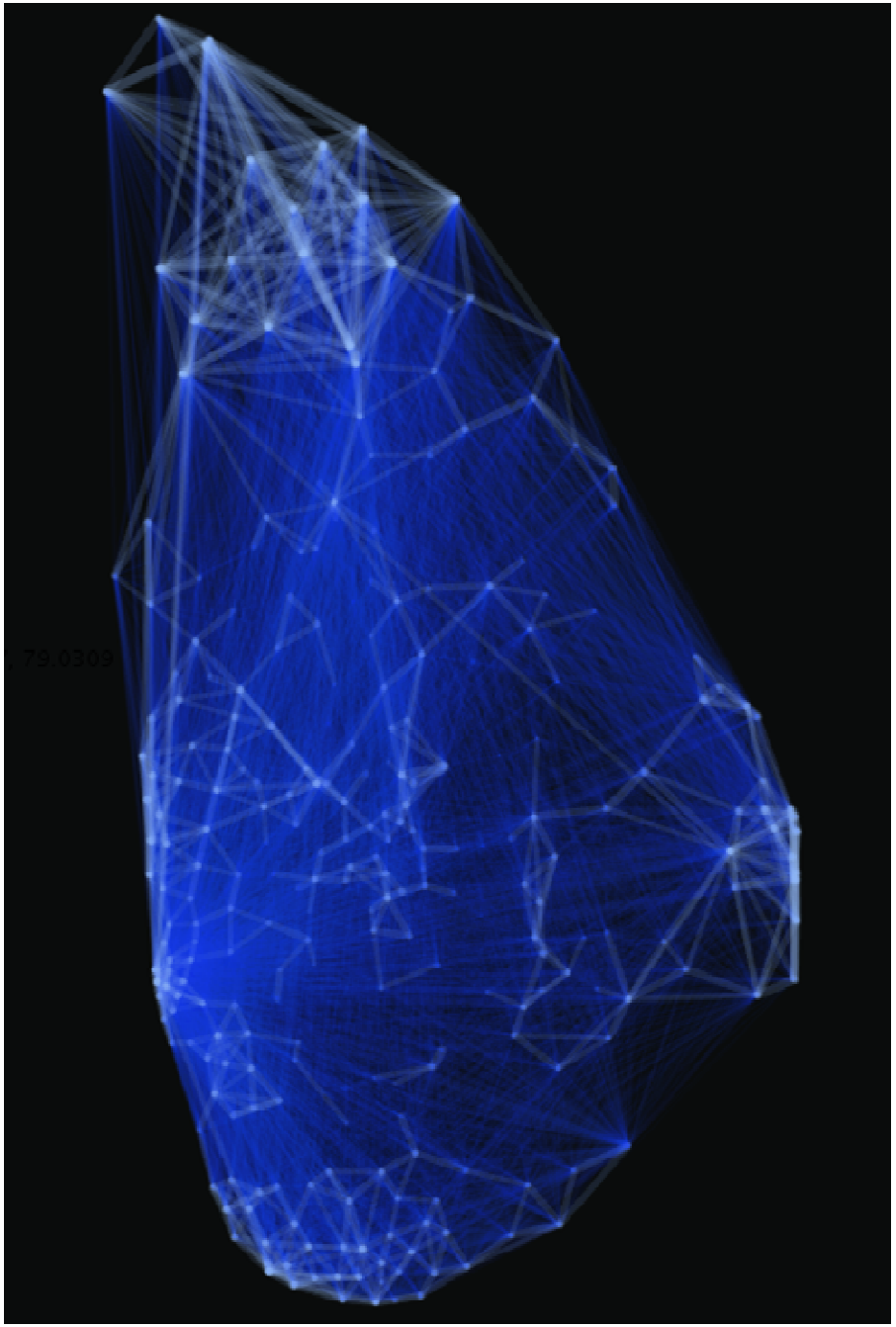


- Cluster 1 corresponds to more commercial area
- Cluster 2 corresponds to more residential (and mixed-used regions)

We can understand the geospatial distribution of the social networks in the country

$$Normalized\ calls\ (DSD_1, DSD_2) = \frac{No.of\ calls\ (DSD_1, DSD_2)}{Population\ (DSD_1) \times Population\ (DSD_2)}$$

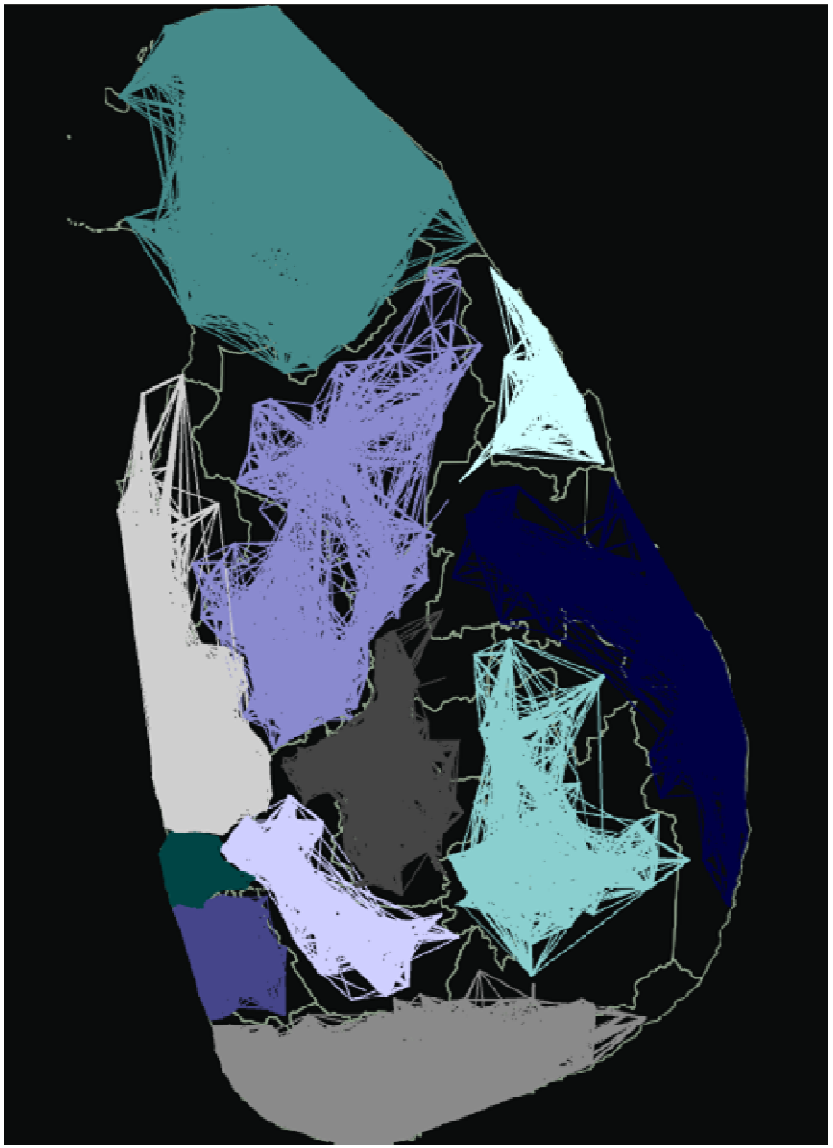
- Each link represents the normalized volume of calls between two DSDs
  - Divisional Secretariat Division (DSD) is a third level administrative division; 331 in total in LK



LIRNEasia
Pro-poor. Pro-market.

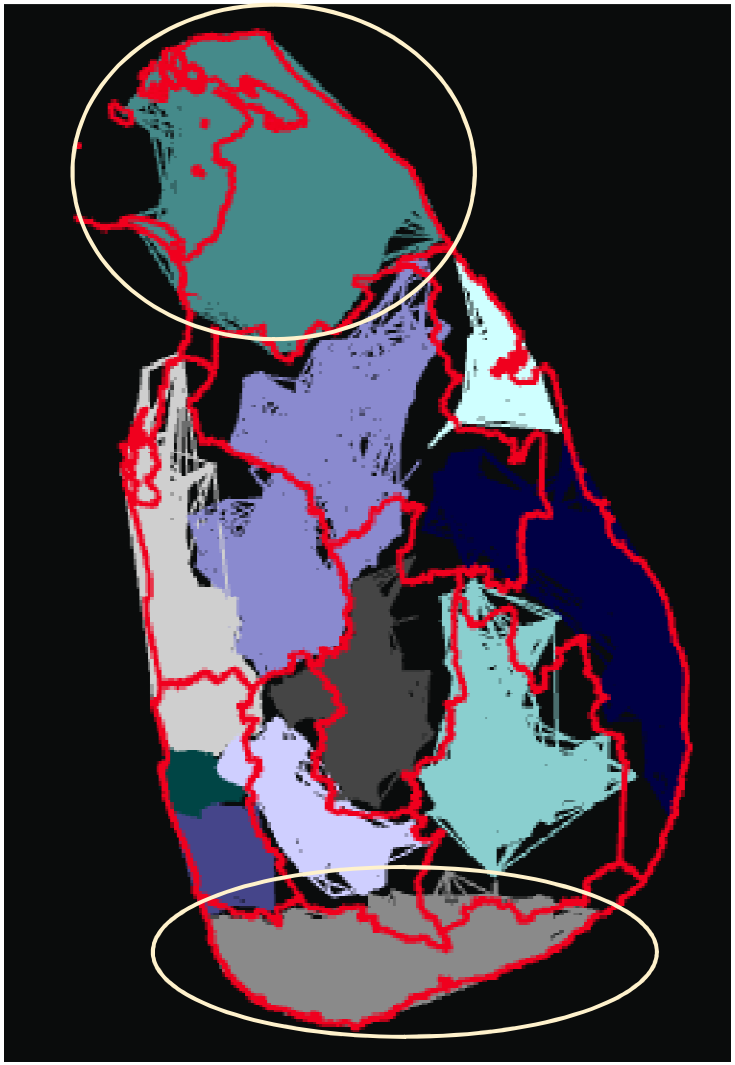
We can understand community structures

- The social network is segregated such that overlapping connections between communities are minimized using a modularity approach
- For Sri Lanka, the optimal number of communities discovered by the algorithm was 11



How much do these communities mesh with existing administrative boundaries?

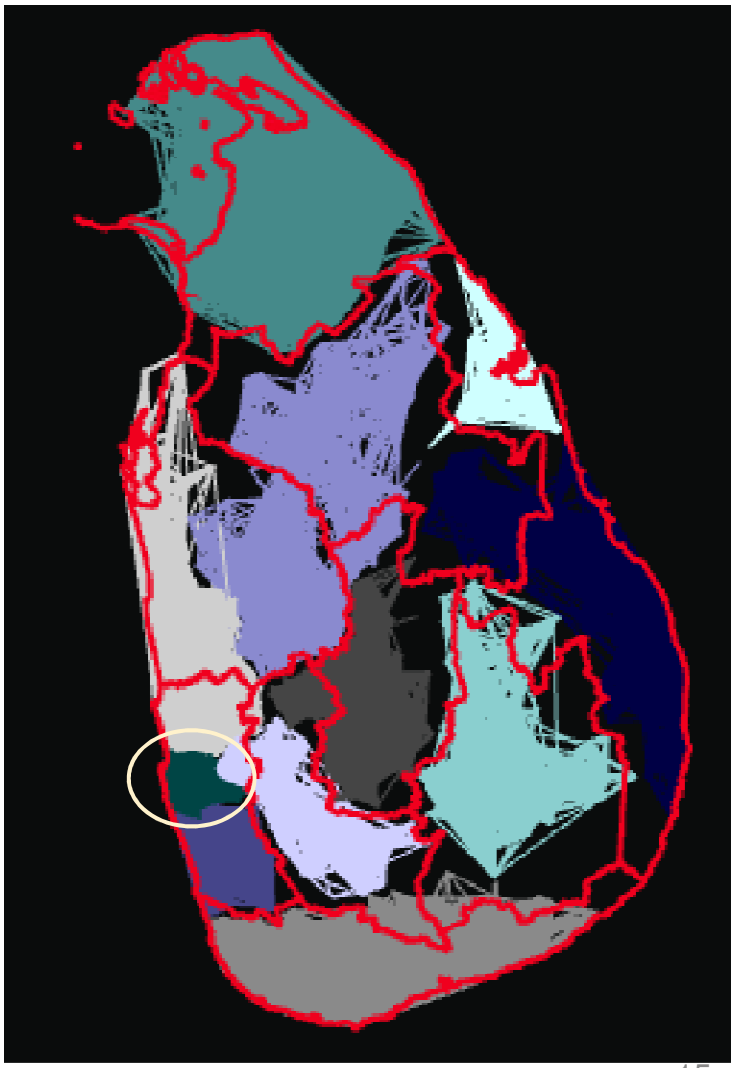
- Southern and Northern provinces have the highest similarity to their respective provincial boundaries.



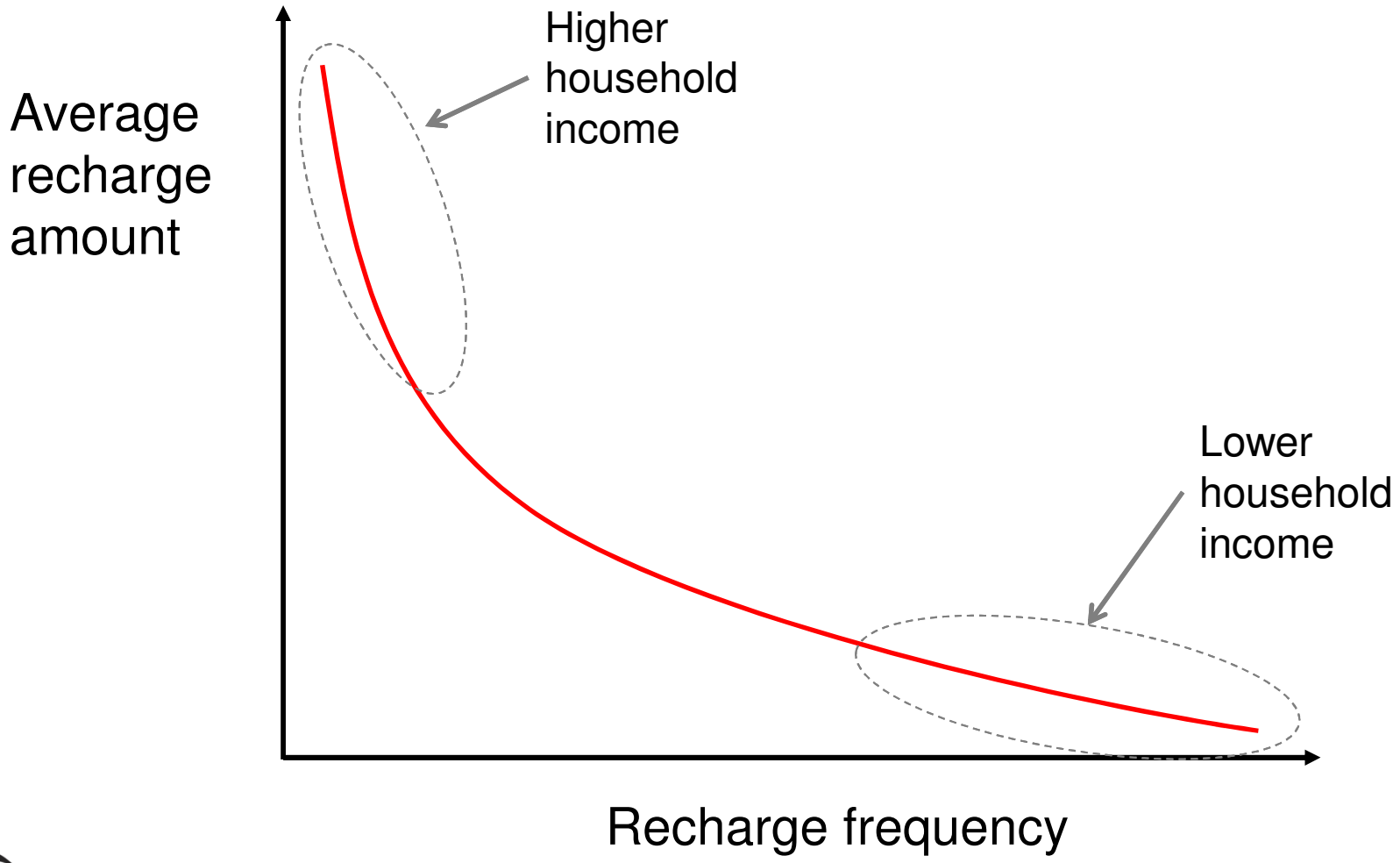
LIRNEasia
Pro-poor. Pro-market.

How much do these communities mesh with existing administrative boundaries (contd.)?

- Colombo district is clustered as a single community and Gampaha is merged with North Western Province



We can leverage the differential patterns of airtime recharge ….

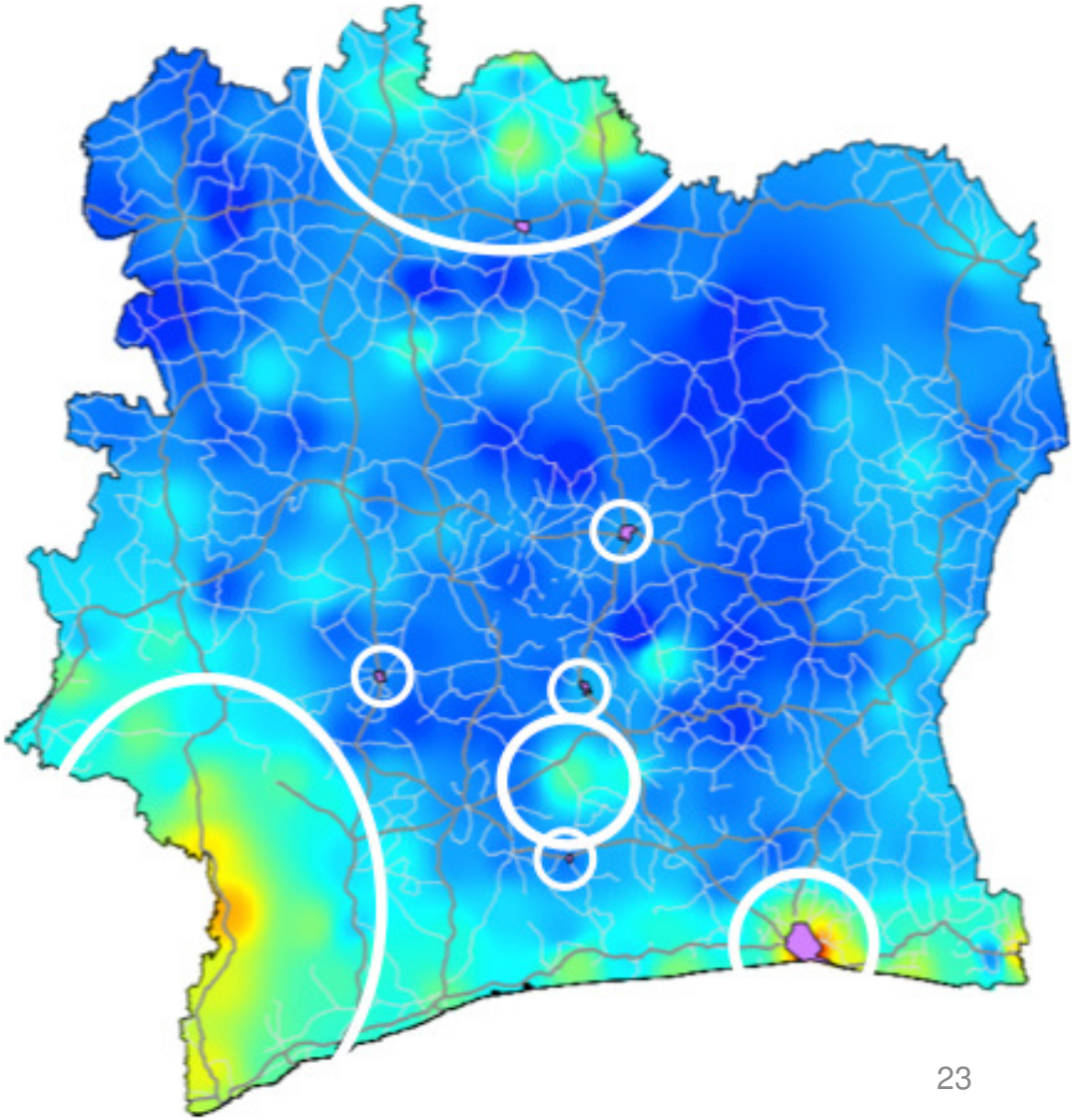


LIRNEasia
Pro-poor. Pro-market.

… to create poverty maps

An example from Cote d'Ivoir

- Source: Thoralf Gutierrez, Gautier Krings, Vincent D. Blondel, (2013). Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. Available at http://arxiv.org/abs/1309.4496

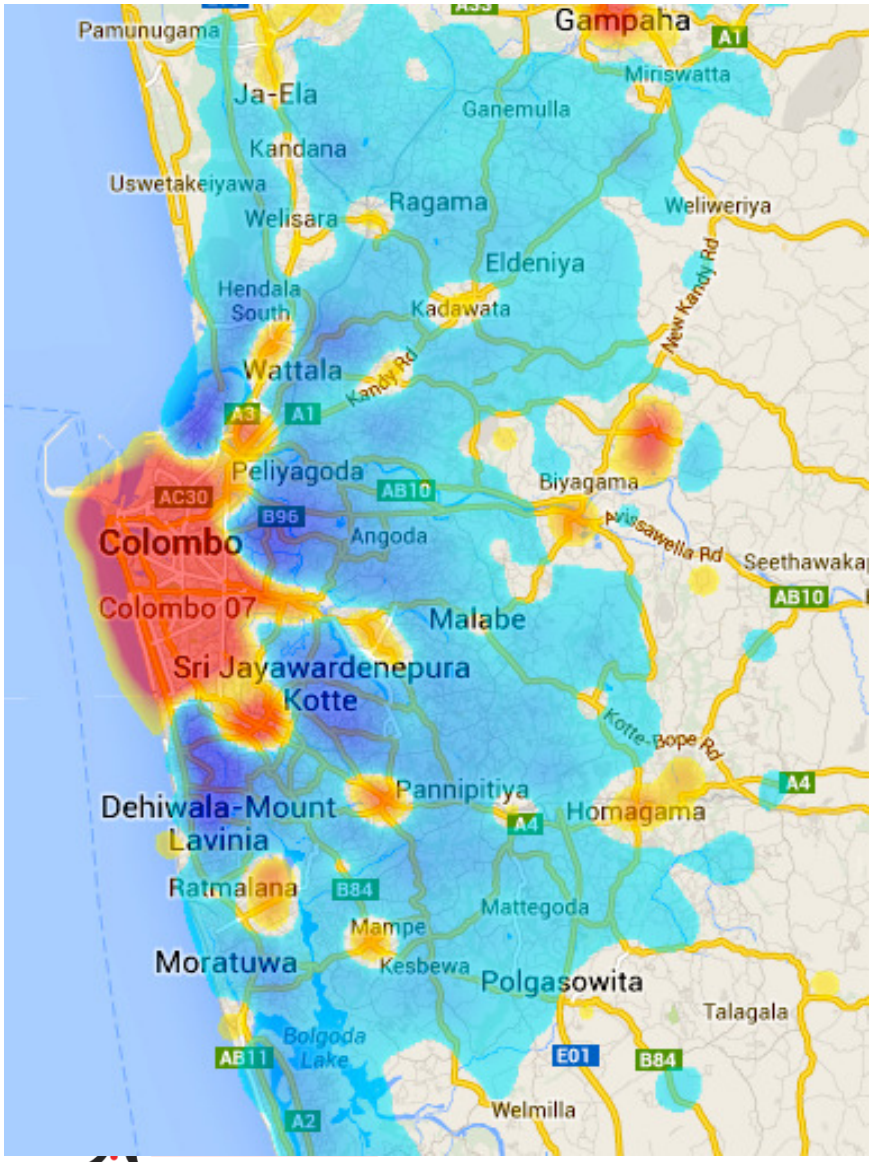


LIRNEasia
Pro-poor. Pro-market.

What's the use of such work?

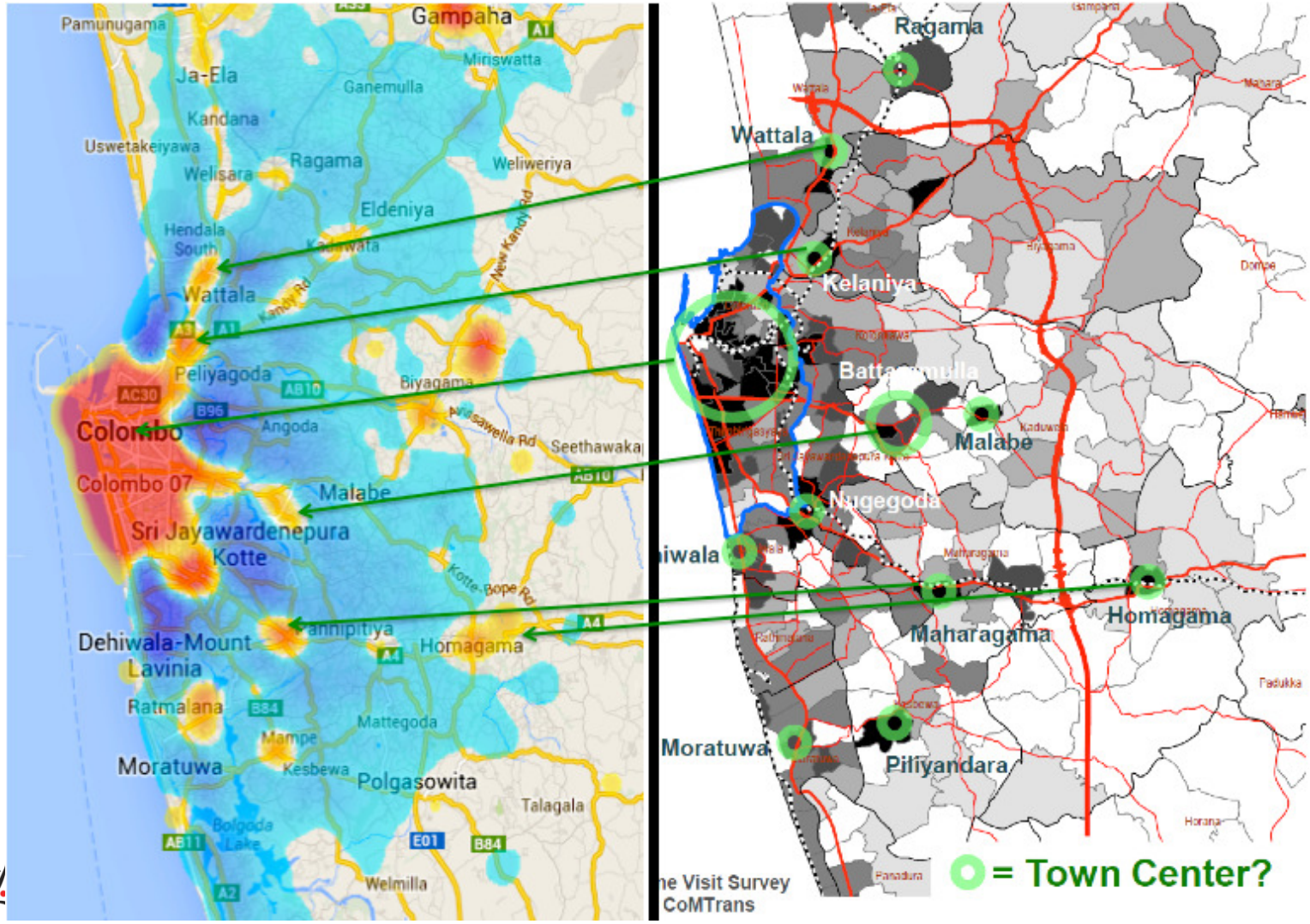
- Can bring timely evidence into the policy making process in developing economies

An example of timely policy-relevant evidence

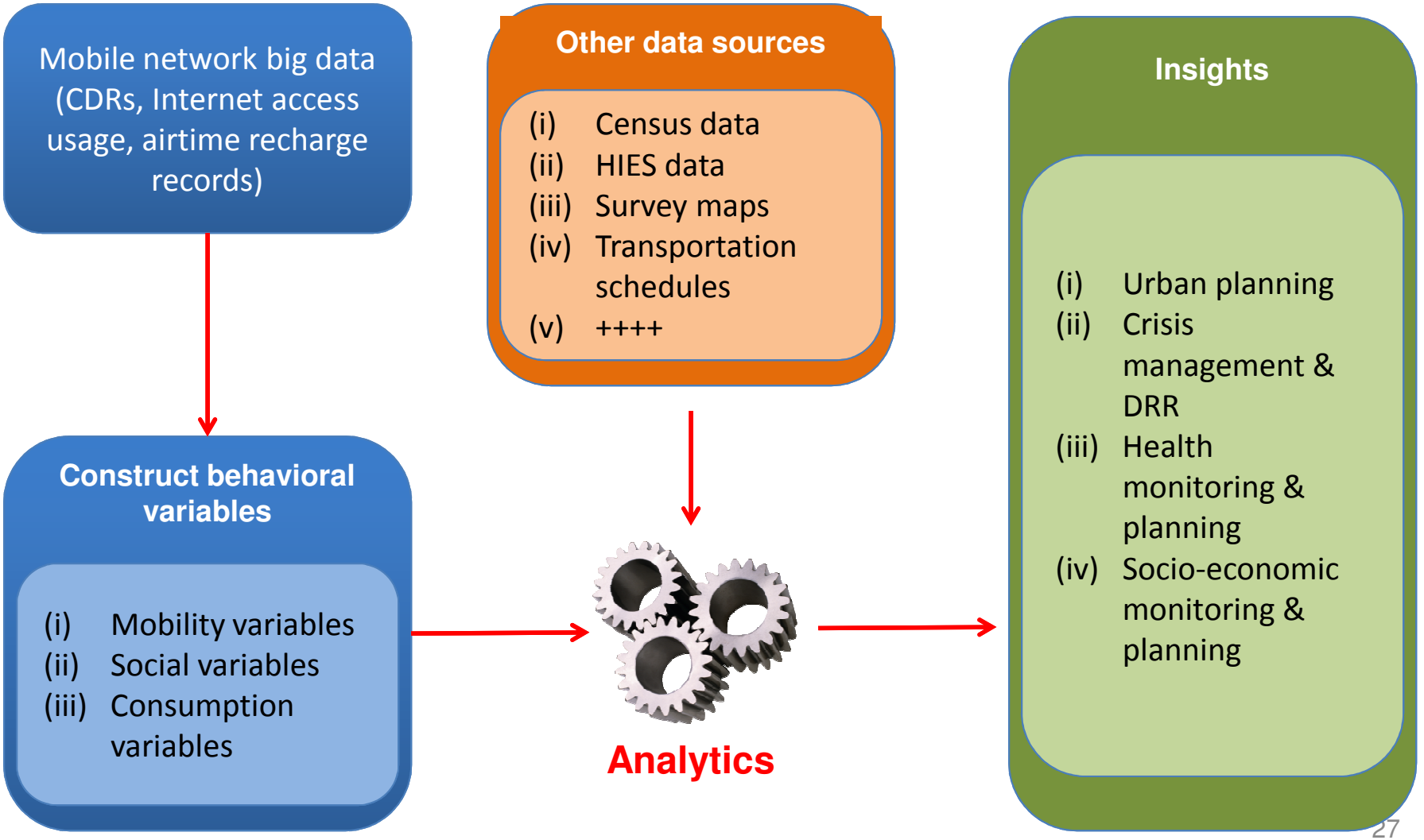


- The image on the left depicts relative density of people in Colombo city and the surrounding regions at 1300 compared to 0000 (midnight the previous day) on a normal weekday.
- The yellow to red colors depict areas whose density has increased relative to midnight. The blue color depicts areas whose density has decreased relative to midnight (the darker the blue, the greater the loss in density). The clear areas are those where the overall density has not changed.

Our findings closely match results from expensive & infrequent transportation surveys

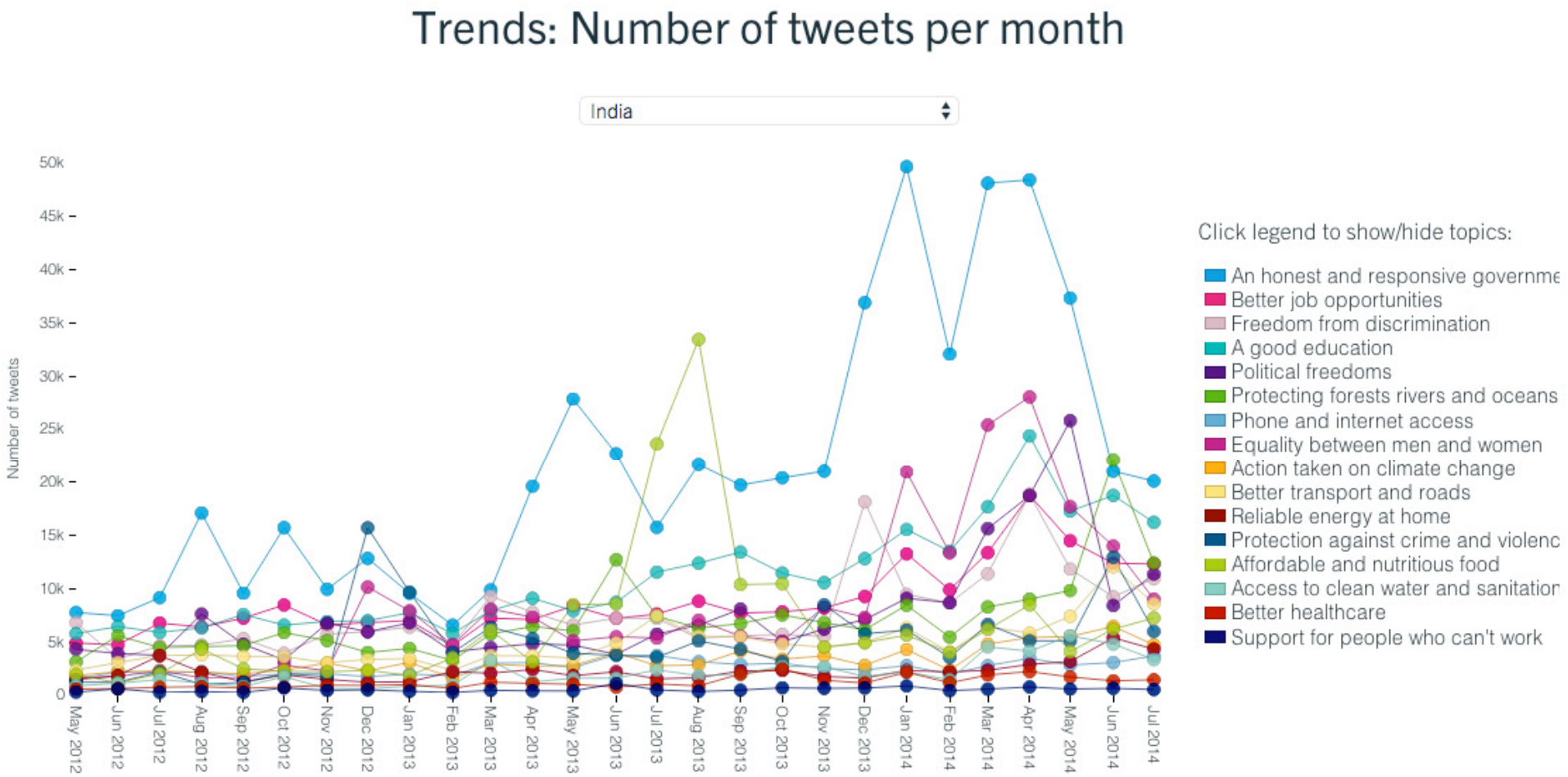


Taking a broader view: the overall process of leveraging mobile network big data for development



What are other potential data sources for big data for development?: Some examples

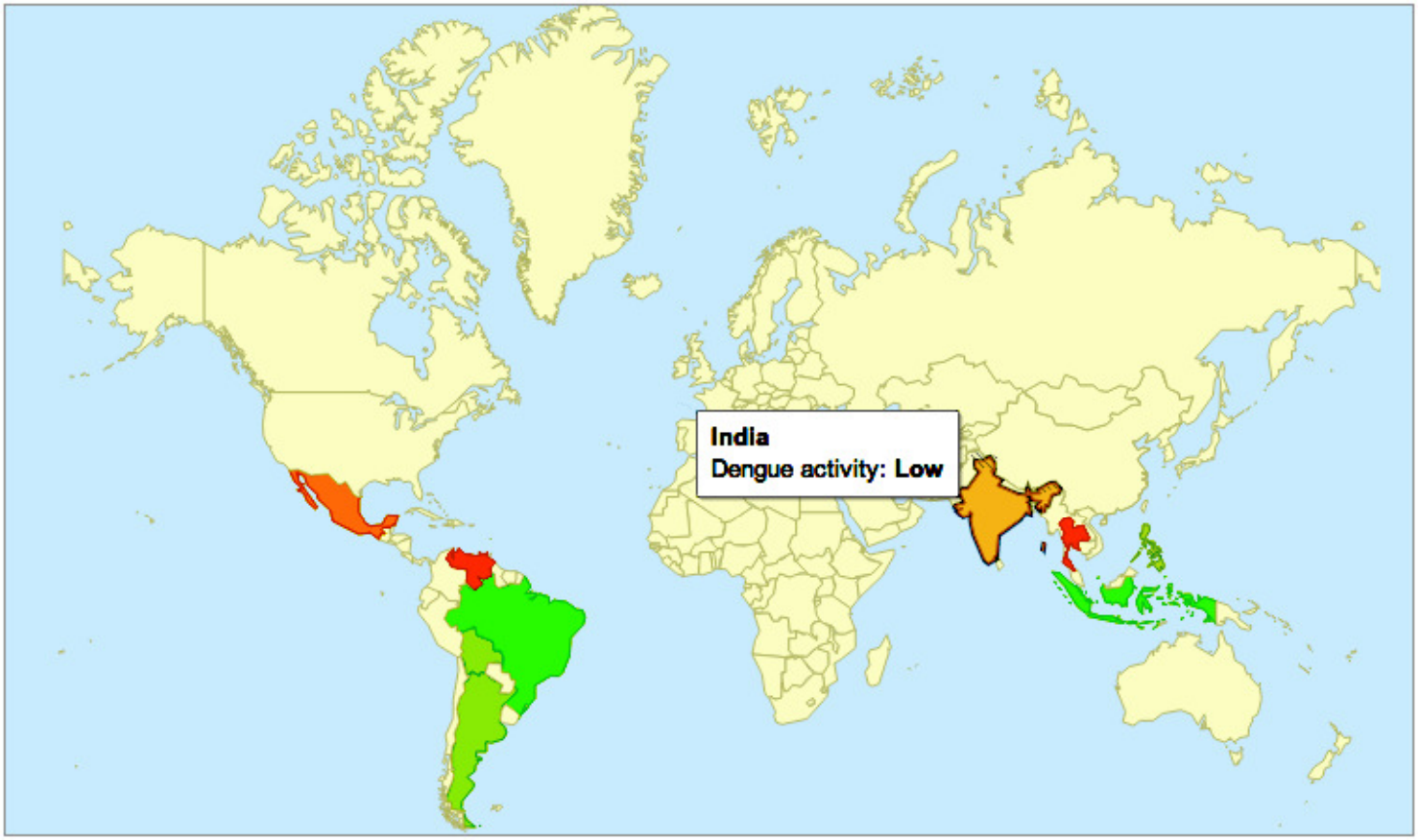
UN Global Pulse Global Post-2015 Twitter Conversation



- http://post2015.unglobalpulse.net/
- Searches approximately 500mn tweets daily across 16 key development topics (represented by about 25k keywords in multiple languages)

LIRNEasia
Pro-poor. Pro-market.

Google Dengue Trends (http://www.google.org/denguetrends/)



- Uses global search activity
- Follow-up to the successful (or not) Google Flu Trends (http://www.google.org/flutrends/), but more on that later

LIRNEasia
Pro-poor. Pro-market.

That's cool, but who can get such work done?

- **Answer: collaborative inter-disciplinary teams**
  - Data mining / computer science
  - Information science
  - Econometrics/ statistics
  - Social science (economics, political science, sociology, anthropology, etc.)
  - Domain expertise (e.g. urban & transport policy, etc.)

Spillovers from the big data phenomenon

- Greater emphases on data driven decision making
- Better tools for data extraction, analysis and visualization even for 'small data'
  - PDF to excel convertors e.g. Tabula
    - http://tabula.nerdpower.org/)
  - Google Fusion Tables
    - https://support.google.com/fusiontables/answer/2571232
  - Data Driven Documents (D3)
    - http://www.D3js.org

Analytical challenges of working with big data

- Data provenance & data cleaning
- Representativeness
- Behavioral change
- Ground context
- Causation vs. correlation
- The role of ‘small data’ for verification as well as for bootstrapping
- Transparency & replicability

Data provenance

- ***Data provenance*** is the process of tracing the pathways taken by data from the originating source through all the processes that may have mutated and/or replicated the data up until it is utilized for the analyses in question
- How relevant is it for big data analyses?
  - Quite relevant though it may not be feasible to establish provenance to the extent desired by the scientific community
  - E.g. when analyzing CDRs, some operators may have multiple records for forwarded calls. Not knowing this, can potentially affect social network analysis

Data cleaning

- Same steps apply for big data as for 'small data':
  - First, verify that quantitative and categorical variables are coded as it should be
  - Second, remove outliers.
- The methods though might have to be different
  - E.g. when removing outliers, you have to use decision tree algorithms or heuristics such as 3-sigma from the mean, etc.

Is the data representative?

- Large volume may make sampling rate irrelevant but doesn't necessarily make the data representative
  - Question: can you think of some representativeness issues in reference to mobile network big data?

Changes in behavior

- Once you know the process people will try to beat it
  - E.g. Google page rank and Search Engine Optimization (SEO)
- Digitized online behavior can be subject to self-censorship and the creation of multiple personas
  - E.g. your Facebook and Twitter posts can mirror how narcissistic you are
    - http://www.dailymail.co.uk/sciencetech/article-2340594/University-study-finds-Facebook-Twitter-fuel-narcissism-different-ways.html
- Transactional data (by-product of doing something else e.g. making phone calls) is therefore the best form of behavioral data
  - But even transactional data can be subject to behavioral changes
  - Question: can you think of an example with respect to mobile network big data?

Ground context

- Knowing and understanding real world context is important, otherwise you might make false inferences
  - Example from Sri Lanka:
    - The majority of airtime reloads occur through scratch cards.
    - Higher denomination cards are not easily available
    - Question: Based on what you learnt earlier about how to differentiate between those form lower and higher income households, can you identify a potential problem when trying to develop poverty maps from mobile network big data?

Causation versus Correlation

- 'There are often more police in precincts with high crime, but that does not imply that increasing the number of police in a precinct would increase crime'
  - Source: Varian, H. R. (2013). Big Data: New Tricks for Econometrics. Available at http://people.ischool.berkeley.edu/~hal/Papers/2013/ml.pdf
- Big Data analyses can **ONLY** reveal correlation **NOT** causation
- You can get to causal effect in a Big Data world by running experiments, but you and I cannot do that easily

But in some instances correlation can be enough to take actions

- E.g. Street Bump mApp in Boston
  - http://www.cityofboston.gov/doit/apps/streetbump.asp
  - Released by Boston City Hall for smartphones
  - Once loaded on smartphones, uses the phone's accelerometer to detect potholes when users are driving, and then notifies City Hall of location.
  - Certain changes in accelerometer readings correlates with passing over a pothole, but no guarantee of causal effect.
  - Question: Why is correlation enough to take action in this case?

The role of traditional “small data” in verification and bootstrapping

- Verification:
  - E.g. we need transport survey data to verify if the patterns of mobility discovered using mobile network big data meshes with real conditions
- Bootstrapping
  - E.g. We can use mobile network big data from around the time of the census to build models to approximate socio-economic levels from just mobile network big data. This is then used to reverse engineer approximate census/ poverty maps in future when other data is not easily available.
  - See Frias-Martinez, V., & Virseda, J. (2012). On the relationship between socio-economic factors and cell phone usage.
- In both cases, big data is complement and not a substitute for ‘small data,’ except in some cases.

Transparency & replicability

- A lot of interesting big data sources are held by private entities
- The benefits of peer-review not yet easily available for much big data based research
- Less transparency and replicability means less chance of catching errors and improving the models

An illustration of the challenges using the example of Google Flu Trends (GFT)

- Since being launched successfully in 2008, subsequent research by others has found problems with how well GFT can predict incidence of Flu
- Problems:
  - Original results based on bootstrapping with data from CDC, but there was no subsequent fine-tuning
  - Over time more and more people started using Google to search for information on health issues, creating rebound effects post introduction of GFT
    - Influenza-like illness rates did not necessarily correlate with actual influenza virus rates
  - Original research article did not reveal the specific 45 search terms that were used to build the model

Analytical challenges of working with big data

- Data provenance & data cleaning
- Representativeness
- Behavioral change
- Ground context
- Causation vs. correlation
- The role of 'small data' for verification as well as for bootstrapping
- Transparency & replicability

For more discussion see my guest blog post at
http://www.unglobalpulse.org/analytical-challenges-big-data

Other challenges of trying to leverage big data for development

- Getting access to data
- Privacy

Getting access

- Some, such as Twitter data are accessible to all (with some volume limits)
- But for most private data sources, initially it will be through ad hoc means, though that is changing
  - E.g. Orange Data for Development competition & Telecom Italia Big Data challenge
    - See http://www.d4d.orange.com/en/home & http://www.telecomitalia.com/tit/en/bigdatachallenge/contest.html
  - LIRNEasia and others working towards opening up the playing field for others

Privacy

- There are personal and collective privacy implications
- The difficulty is that:
  - Informed consent is meaningless in a big data world
  - People actually don't really know what their generalizable privacy needs or how their preferences might evolve
- Ways to deal with the privacy issue:
  - Anonymization (with caveats)
  - Different levels of disaggregation of shared data
  - Evolution from a rights based approach to a harms based approach
- However more research still required

What you should have gained from today's lecture (hopefully ☺)

- Some understanding of what big data means
- An appreciation for some of the developmental opportunities from leveraging big data using different sources
- A brief overview of the spillover effects of the big data phenomenon
- An appreciation of some of the analytical and other challenges concerning the use of big data for development

Thank you