Big Data for Development (BD4D)

Sriganesh Lokanathan, LIRNE*asia*

LIRNE*asia* Research Planning Meeting
20-21 December 2012
Colombo

LIRNE*asia*
www.lirneasia.net

This work was carried out with the aid of a grant from the International Development Research Centre, Canada a

IDRC CRDI

Canada

Agenda

- What is Big Data?
- Big Data for Development (BD4D)
- Examples of BD4D using telecom data
- LIRNE*asia*'s exploratory work in 2012-2014
- The Sri Lankan telecom BD datasets!
- The process of converting telecom BD into insights
- Research questions
- Challenges

What is Big Data?

- Big Data is a popular but ***subjective*** term that describes large Volumes of high Velocity, complex and Variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.
- Key characteristics
  - Data is often transactional (i.e. TGI) captured as a by-product of doing things (such as providing telephone service, processing payments, etc.) but not necessarily (e.g. social media, sensor readings)
  - Facilitated by the vast drops in the cost of storing and retrieving information
  - Facilitated by exponential growth in computer power and memory (data can reside in persistent memory instead of disk and tape)
  - Facilitated by vast improvements in techniques for performing machine learning and reasoning

Big data spans 4 dimensions

- Volume
  - Volume of digital data growing exponentially from transaction-based commercial data, social media data, sensor data, etc.
- Velocity
  - High rate of production of digital data means it requires it must be processed just as fast to meet demand for insights
- Variety
  - Data now comes in a variety of media (text, video, audio) and formats (structured & unstructured).
  - More importantly the datasets themselves are varied and across different domains, leading to vast possibilities in analyses of correlations and new insights
- Veracity
  - With the data deluge, the occurrence of imprecise, uncertain or faulty data is also increasing. The Velocity and Variety dimensions positively correlate to increased need for veracity.

Agenda

- What is Big Data?
- Big Data for Development (BD4D)
- Examples of BD4D using telecom data
- LIRNE*asia*'s exploratory work in 2012-2014
- The Sri Lankan telecom BD datasets!
- The process of converting telecom BD into insights
- Research questions
- Challenges

BD4D: taking evidenced-based policy making to the next level

- In our hyper-connected world, there are streams of data being generated (public and private) that offer possibilities for ***rich, real-time*** insights especially when mixed and mashed

- Big data vs. traditional surveys
  - Can find out what people actually did versus what they recalled
  - Real-time insights are possible as opposed to the time lag from traditional surveys

Sources of BD4D

Source	Comprehensive coverage of the poor?
Social media (e.g. Twitter)	?
Online queries (e.g. Google flu trends or more recently Google dengue trends)	?
Telecom network operator data: • Call Detail Records (CDRs) • Top-up histories	✓

Agenda

- What is Big Data?
- Big Data for Development (BD4D)
- Examples of BD4D using telecom data
- LIRNE*asia*'s exploratory work in 2012-2014
- The Sri Lankan telecom BD datasets!
- The process of converting telecom BD into insights
- Research questions
- Challenges

Examples of BD4D using telecom data

- Amy Wesolowski and Nathan Eagle’s work in Kenya (mainly Kibera slums)
  - Migration patterns
  - Understanding the effect of human mobility patterns in the spread of malaria
  - Understanding how reload patterns in a region could serve as a” smoke signal” of economic shocks
- Vanessa Frias-Martinez’s (Telefónica Research) work on socio-economic mapping
  - Has demonstrated that calling patterns can be used to identify the socioeconomic level of a population, which in turn may be used to infer its access to housing, education, healthcare, and basic services such as water and electricity

Examples of BD4D using telecom data contd.

- Linus Bengtsson and Xin Liu (Karolinska Institutet in Sweden) work on population displacements after 2010 Haiti earthquake
  - Used data from Digicel, Haiti's largest cell phone provider, to determine the movement of displaced populations after the earthquake, aiding the distribution of resources
  - Modeled spread of Cholera predicting future hotspots in the country based on population movements after the earthquake.
- MIT research on early warnings of personal illness
  - Found students who came down with fever/ flu moved less and made fewer calls in the mornings and late evenings. Software training then could identify occurrence of illness in individuals with 90% accuracy

Agenda

- What is Big Data?
- Big Data for Development (BD4D)
- Examples of BD4D using telecom data
- LIRNE*asia*'s exploratory work in 2012-2014
- The Sri Lankan telecom BD datasets!
- The process of converting telecom BD into insights
- Research questions
- Challenges

BD4D – LIRNE*asia* exploratory work in 2012-2014

- LIRNEasia has negotiated telecom network data from two operators in Sri Lanka
- Over the course of the two years, we are:
  - Conducting exploratory research on answering a ***few*** social science questions
  - Developing a framework with privacy and self-regulatory guidelines for the collection, use and sharing of mobile phone data. This would be a consultative process amongst relevant stakeholders such as operators, research organizations as well as relevant government agencies
- Partners:
  - Auton Lab (Carnegie Mellon University) will provide technical and analytical support

LIRNEasia
www.lirneasia.net

Tool(s)

- Mainly T-cube (tcube.autonlab.org) from Auton Lab
  - With customizations (for visualizations plus specific analytical modules)
- R (statistical software) + visualization modules

T-cube (tcube.autonlab.org)

- T-Cube stores data models in computer memory and performs rapid retrieval for advanced analytics and interactive mining of large data sets.
- T-Cube facilitates quick and reliable discovery of complex patterns where such tasks were previously considered infeasible.
- Amongst others, it has been used to detect patterns of foodborne illness in the US food supply, discover emerging trends in equipment maintenance, and in finding unexpected increases in human disease rates across the globe.
- LIRNE*asia* has employed the software in the past in its Real-Time Bio-surveillance Project

Agenda

- What is Big Data?
- Big Data for Development (BD4D)
- Examples of BD4D using telecom data
- LIRNE*asia*'s exploratory work in 2012-2014
- The Sri Lankan telecom BD datasets!
- The process of converting telecom BD into insights
- Research questions

The datasets

- Datasets from two operators (Company A and Company B):
  - CDRs (record of incoming and outgoing calls and SMS-es)
  - Internet access records
  - Top-up records
- Datasets only contain annonymized numbers and no actual phone numbers
- We are getting for the period 1st April 2012 to 15 October 2012 (6.5 months)
  - Covers some important dates: Sinhala and Tamil New Year (April), Tsunami warning (11th April), Vesak (May), T20 World Cup

Dataset 1: Call Detail Record (CDR)

- Flag to identify type of record: if its an incoming or outgoing call, terminating or originating SMS
- Antenna/ Cell ID + Antenna
- IMEI number
  - Allows us to identify the type of handset
- Annonymized subscriber number
- Annonymized other number
- Country code of other number when it is an international number
  - Derived and not part of core CDR
- Date & time of call/ SMS
- Duration of call

Dataset 2: Internet Access Record (IAR)

- Antenna/ Cell ID + antenna
- IMEI number
  - Allows to identify the type of handset
- Annonymized subscriber number
- Date & time of access
- Duration of internet access
- Flag to denote if it was a 2G or 3G connection
- Access Point Name (APN)
  - A rough method to differentiate between internet access from low end phones, high end phones and dongles
- Uplink and downlink data volumes for a session

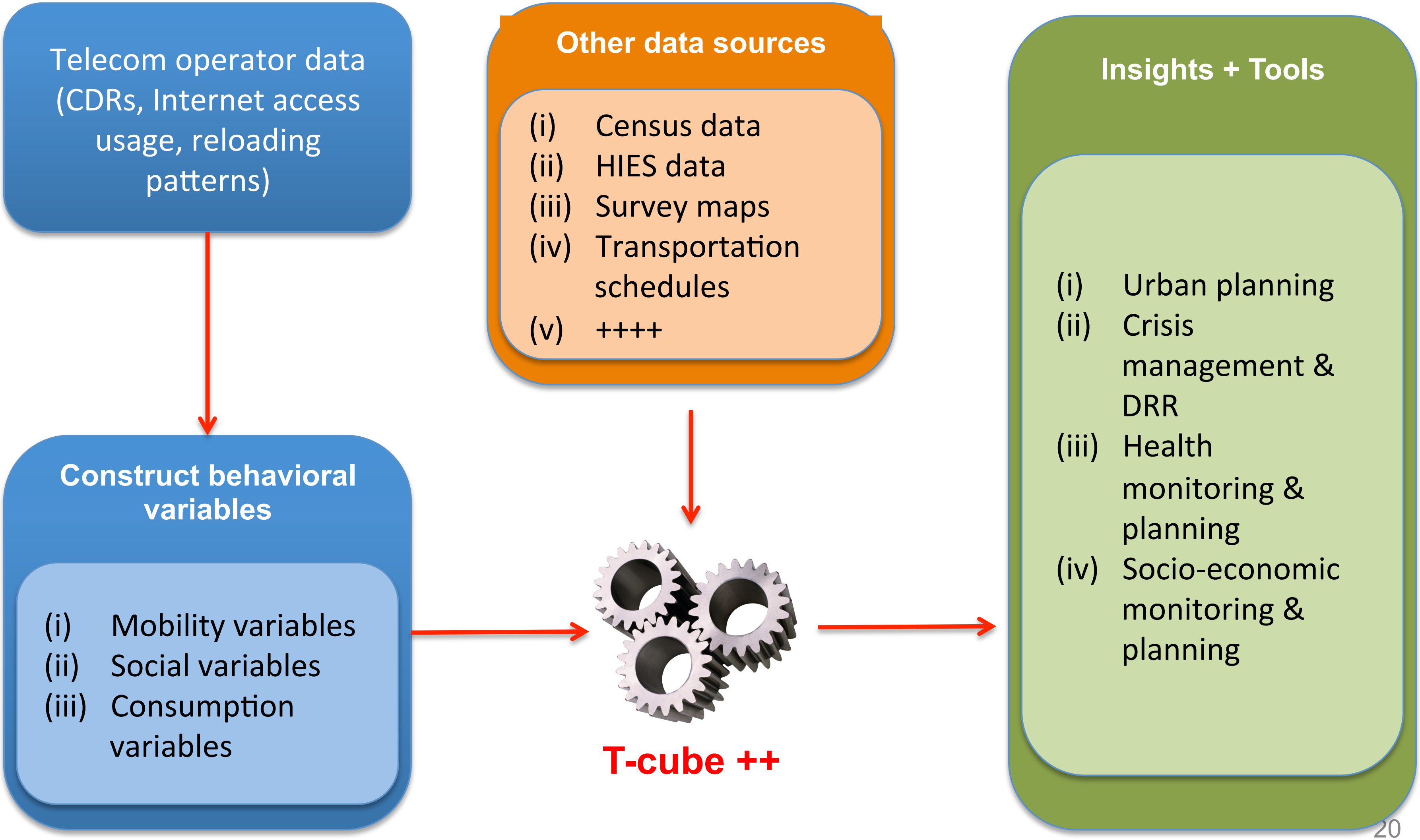
Dataset 3: Top-up records

- Flag to denote whether reload was via load transfer (electronic top-up) or via scratch card.
- Annonymized subscriber number
- Top-up amount
- Phone balance after top-up
- Date and time of top-up

Agenda

- What is Big Data?
- Big Data for Development (BD4D)
- Examples of BD4D using telecom data
- LIRNE*asia*'s exploratory work in 2012-2014
- The Sri Lankan telecom BD datasets!
- The process of converting telecom BD into insights
- Research questions
- Challenges

BD4D: the overall process



Telecom operator data (CDRs, Internet access usage, reloading patterns)

Other data sources

(i) Census data
(ii) HIES data
(iii) Survey maps
(iv) Transportation schedules
(v) ++++

Insights + Tools

(i) Urban planning
(ii) Crisis management & DRR
(iii) Health monitoring & planning
(iv) Socio-economic monitoring & planning

Construct behavioral variables

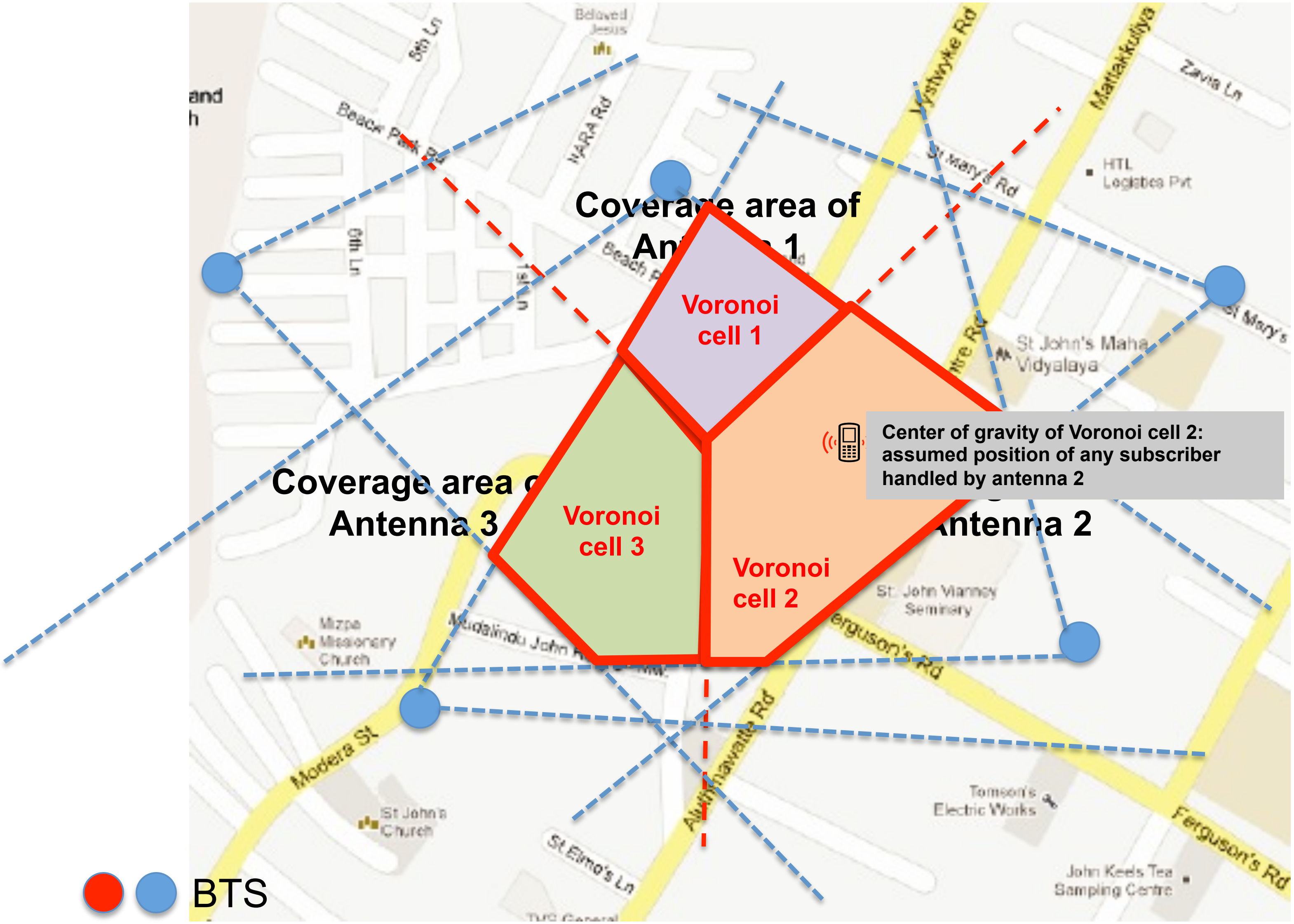
(i) Mobility variables
(ii) Social variables
(iii) Consumption variables

T-cube ++

Behavioral variables from telecom BD

- Mobility variables
  - Diameter of mobility and social network
  - Radius of gyration
  - Mobility profiles/ patterns
- Social variables
  - Connections (***mutality/ reciprocity***, network closure, ***propinquity***)
  - Distributions (bridges, centrality, density, distance, structural holes, tie strength)
  - Segmentation (community detection, social circles, clustering coefficient, cohesion)
- Consumption variables
  - Number and duration of usage (call, missed-call, SMS/MMS, internet)
  - Size and frequency of top-ups
  - Handset type and features

How to approximate a person's position from a CDR's Cell ID



Center of gravity of Voronoi cell 2: assumed position of any subscriber handled by antenna 2

Limitation of location/ mobility estimates from CDR data

- Voronoi cell area (and thus potential user location) depends on BTS density
  - Small cell in urban areas (high BTS density)
  - Larger cell in rural areas (low BTS density)
- Data can be sparse when used for mobility
  - Can find a location only when a person made or received a call/ sms, started an internet connection or topped-up
  - Building a mobility profile requires a larger time frame and route calculations than when we have cell-hand off data

How do we know who the poor (BOP & BOP MEs)

- How to identify the BOP?
  - Our T@BOP surveys showed that the poor top-up in small amounts but with higher frequency
  - Utilize reciprocity index from CDRs
    - studies have shown lower socio-economic individuals have a lower reciprocity index
- What about BOP MEs?
  - Get informed consent from ME survey respondents in Sri Lankan sample to look at their call patterns (if they utilize Company A or B's mobile service)
  - Find patterns through machine learning of known set to identify others

Agenda

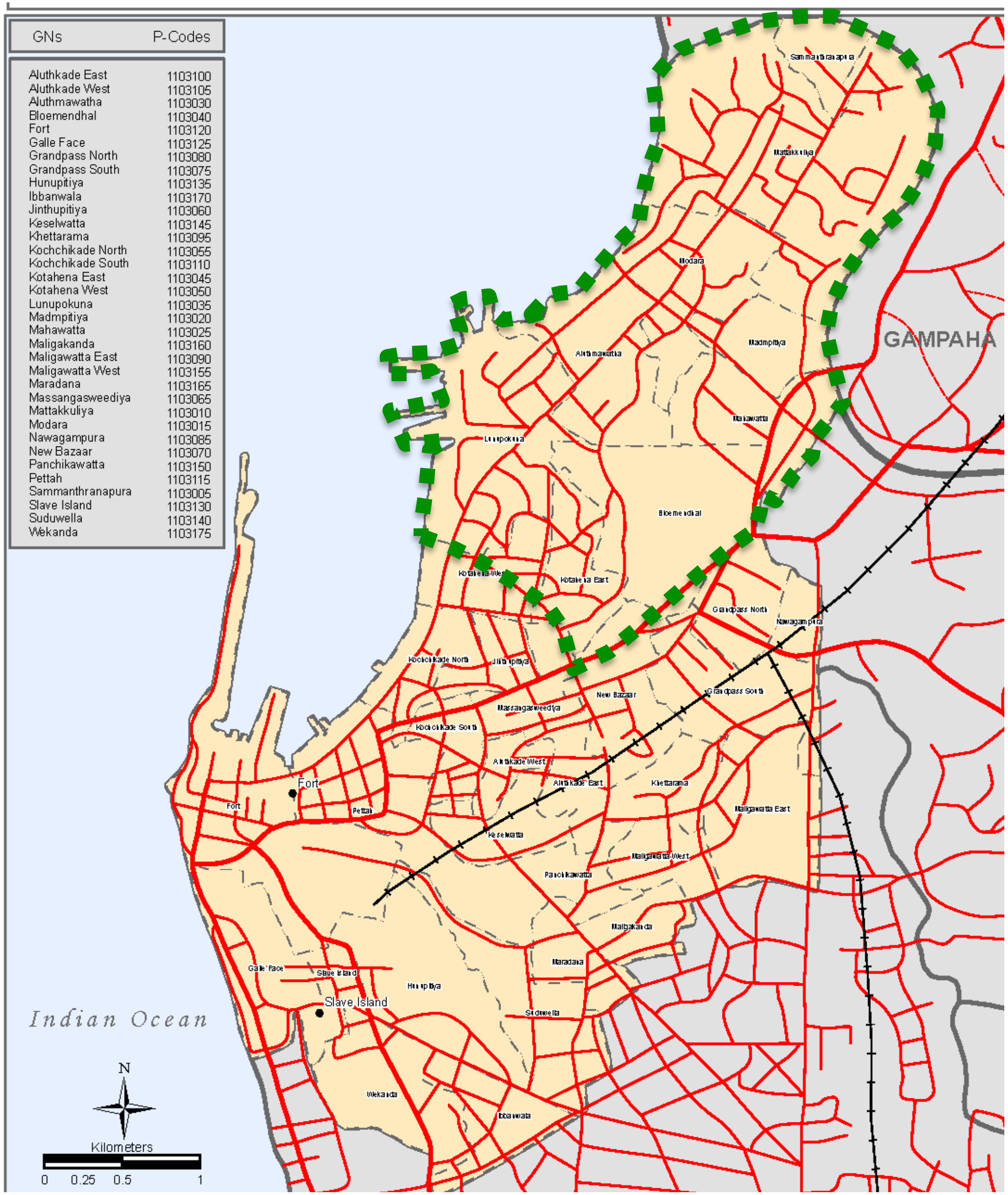
- What is Big Data?
- Big Data for Development (BD4D)
- Examples of BD4D using telecom data
- LIRNE*asia*'s exploratory work in 2012-2014
- The Sri Lankan telecom BD datasets!
- The process of converting telecom BD into insights
- Research questions
- Challenges

North Colombo (Colombo 15 and parts of 13 & 14)

Part of Colombo Division of Colombo District

Grama Niladhari Divisions under North Colombo:

1. Sammanthranapura
2. Mattakkuliya
3. Modara
4. Madmpitiya
5. Mahawatta
6. Aluthmawatha
7. Lunupokuna
8. Bloemendhal
9. Kotahena East
10. Part of Kotahena West



GNs	P-Codes
Aluthkade East	1103100
Aluthkade West	1103105
Aluthmawatha	1103030
Bloemendhal	1103040
Fort	1103120
Galle Face	1103125
Grandpass North	1103080
Grandpass South	1103075
Hunupitiya	1103135
Ibbanwala	1103170
Jinthupitiya	1103060
Keselwatta	1103145
Khettarama	1103095
Kochchikade North	1103055
Kochchikade South	1103110
Kotahena East	1103045
Kotahena West	1103050
Lunupokuna	1103035
Madmpitiya	1103020
Mahawatta	1103025
Maligakanda	1103160
Maligawatta East	1103090
Maligawatta West	1103155
Maradana	1103165
Massangasweediya	1103065
Mattakkuliya	1103010
Modara	1103015
Nawagampura	1103085
New Bazaar	1103070
Panchikawatta	1103150
Pettah	1103115
Sammanthranapura	1103005
Slave Island	1103130
Suduwella	1103140
Wekanda	1103175

LIRNEasia
www.lirneasia.net

Possible research questions/ ideas (1)

Idea	Literature review	Notes
What is the linkage between the level of communication with another city and its corresponding population? • Does communication grow on a super-linear path to city population	The level of communication can say something about productivity, innovative capacity, etc. Andris et al (2009) found that communication is proportional to (population)^1.5 in US cities, hence cities are just scaled versions of one another	Additional items required: • Detailed population data

Possible research questions/ ideas (2)

Idea	Literature review	Notes
How does the population density in North Colombo vary over time: • Intra-day variations • Daily variations over a week • Seasonal changes	Calabrese and Ratti (2006) and Reades et al (2007) papers on Rome using Erlang data to understand patterns of human movement though variations in population density over time	Additional items required: • Detailed population data

Possible research questions/ ideas (3)

Idea	Literature review	Notes
• What is the linkage between a person's socio-economic condition and his/her mobility? • When compared to average urban dweller, do the urban poor travel further or less or is it the same	Frias-Martinez et al (2012) reverse-engineered a Census Map that could demarcate geographic areas by the socio-economic wellbeing of its population.	Additional items required: • Detailed population data • Some measurement of the Socio-economic condition of every household or the very least the percentage breakdown of different classes in a neighborhood.

Possible research questions/ ideas (4)

Idea	Literature review	Notes
- What was the impact of the tsunami warning on 11th April 2012 on population movements	Bengtsson et al (2011) and Lu et al (2012) studies on the impact of the Haiti's earthquake of 2010 on people's movement (amongst other things)	Additional items required: - Detailed population data - Transport maps - Bus and train schedules

Possible research questions/ ideas (5)

Idea	Literature review	Notes
How are the people in North Colombo connected (mobility aspects)? • How do people move in and out of North Colombo. • If they come from outside frequently, where do they come from, how long do the stay • If they migrate in from outside how long do the stay? is the nature of the stay just for work, or just for living (i.e. they live there but work outside) • If they live there, where do they go, how long do they stay there, • How many live and work there • Are the government offices, bill payment locations, tele-centers in the most-optimal locations in comparisons to the movements of the urban poor?	• Wesolowski and Eagle (undated) work on Kiberia slums (refer to Future work section) • Studies by Ratti and Calabreses (multiple) • IBM work in US (Amini, 2011) • Look at Sevtsuk and Ratti (2010) on estimating daily routines/ traffic patterns from mobile phone data. Has a good methodology section with limitations	Additional items required: • Detailed population data • Transport maps • Bus and train schedules • Geo-locations of specific service locations

Possible research questions/ ideas (6)

Idea	Literature review	Notes
How are the people in North Colombo connected (social network aspects)? • How are the social networks of the urban poor different from the average urban resident? Is there social exclusion at the urban poor level • what is the nature of the networks/ connection/ ties to people within the region they live in, the people outside, to the region where they are originally from (if they are "migrant") • What is the frequency of sharing of physical space between them and those they are in contact with	• Refer to Eagle paper on social networks and the algorithms in that paper • Studies by Ratti and Calabreses (multiple)	Additional items required: • Detailed population data

Agenda

- What is Big Data?
- Big Data for Development (BD4D)
- Examples of BD4D using telecom data
- LIRNE*asia*'s exploratory work in 2012-2014
- The Sri Lankan telecom BD datasets!
- The process of converting telecom BD into insights
- Research questions
- Challenges

Challenges

- Mixing datasets from both operators may not be possible due to limits in our negotiated access
  - How do we account for subscriber bias for one network over the other?
  - How do we generalize results to the larger population?
- Resolution limitations may not allow us to differentiate between two parallel roads/ pathways in close proximity
  - Except when the roads run in opposite directions
- It's a long winded process to to get micro-level census data (to GN division)!
- Initially, we will have limited hardware infrastructure, which means we will not always be able to analyze large time-frames