Mobile Network Big Data for Urban Development

LIRNE*asia*

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

LIRNEasia
Pro-poor. Pro-market.

We are a regional non-profit think tank. Our mission is that of

Catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.

LIRNE*asia*

Pro-poor. Pro-market.

Where we work

Big data work only in Sri Lanka in 2012-15
Is being extended to Bangladesh in 2016-17

LIRNEasia
Pro-poor. Pro-market.

"Smart cities," the new buzz phrase

- IBM has been promoting smart cities and big data since the 2000s

- But two visions
  - New smart cities created on green fields, like South Korea's Songdo, OR
  - Improving the functioning of existing cities

What approach is most appropriate for us?

- Can we make our cities smart, on the cheap?

- Yes, by leveraging transaction-generated data ("data exhaust")
  - Open-source analytics and cheap hardware (ours cost less than USD 30k) make it possible
  - Strengths in relevant skills may even give us an advantage ➔ possible to become a regional/international center of excellence

If we want comprehensive coverage of the population, what are the sources of big data in developing economies at this time?
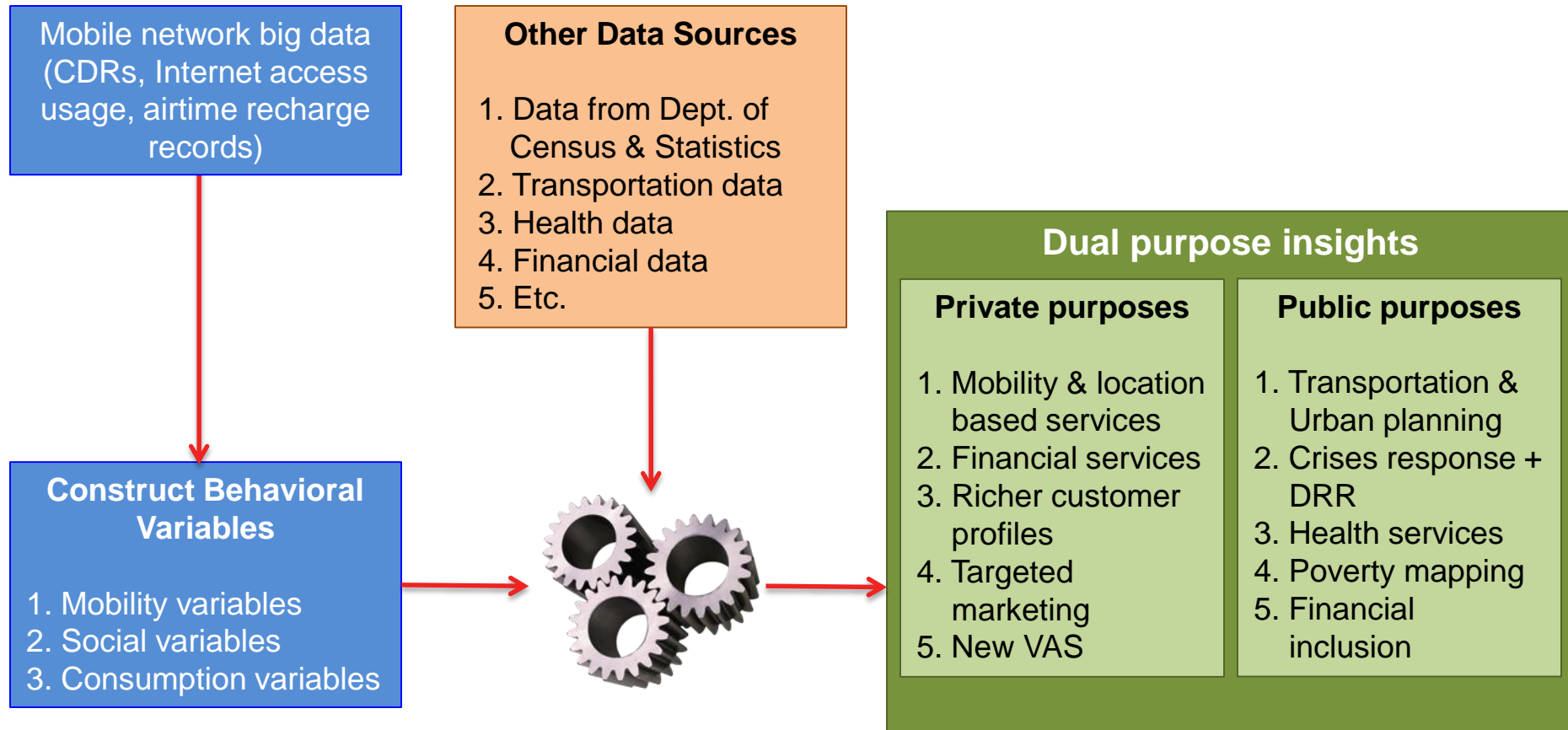
- Administrative data
    - E.g., digitized medical records, insurance records, tax records
- Commercial transactions (transaction-generated data)
    - E.g., Stock exchange data, bank transactions, credit card records, supermarket transactions connected by loyalty card number
- Sensors and tracking devices
    - E.g., road and traffic sensors, climate sensors, equipment & infrastructure sensors, mobile phones communicating with base stations, satellite/ GPS devices
- Online activities/ social media
    - E.g., online search activity, online page views, blogs/ FB/ twitter posts

LIRNE*asia*
Pro-poor. Pro-market.

Mobile Network Big Data is only option at this time

Country	Mobile SIMs/100	Internet users/100	Facebook users/100
	2015	2014	2016
Myanmar	69	2	20.4
Bangladesh	82	10	13.7
Pakistan	67	14	14.3
India	76	18	11.5
Sri Lanka	125	26	19.6
Philippines	120	40	54.6
Indonesia	131	17	33.8
Thailand	125	35	58.9

Sources: https://www.gsmaintelligence.com/;
http://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2015/MISR2015-w5.pdf;
Facebook advertising portal; http://data.worldbank.org/indicator/SP.POP.TOTL

LIRNE*asia*
Pro-poor. Pro-market.

Mobile network big data + other data ➜ rich, timely insights that serve private **as well** as public purposes

**Mobile network big data (CDRs, Internet access usage, airtime recharge records)**

Other Data Sources

1. Data from Dept. of Census & Statistics
2. Transportation data
3. Health data
4. Financial data
5. Etc.

Construct Behavioral Variables

1. Mobility variables
2. Social variables
3. Consumption variables

Dual purpose insights

Private purposes

1. Mobility & location based services
2. Financial services
3. Richer customer profiles
4. Targeted marketing
5. New VAS

Public purposes

1. Transportation & Urban planning
2. Crises response + DRR
3. Health services
4. Poverty mapping
5. Financial inclusion

Data used in the research

- Multiple mobile operators in Sri Lanka have provided four different types of meta-data
  - Call Detail Records (CDRs)
    - Records of calls
    - SMS
    - Internet access
  - Airtime recharge records
- Data sets do not include any Personally Identifiable Information
  - All phone numbers are pseudonymized
  - LIRNE*asia* does not maintain any mappings of identifiers to original phone numbers
- Cover 50-60% of users; very high coverage in Western (where Colombo the capital city in located) & Northern (most affected by civil conflict) Provinces, based on correlation with census data

EXAMPLE: LAND USE MONITORING

Hourly loading of base stations reveals distinct patterns
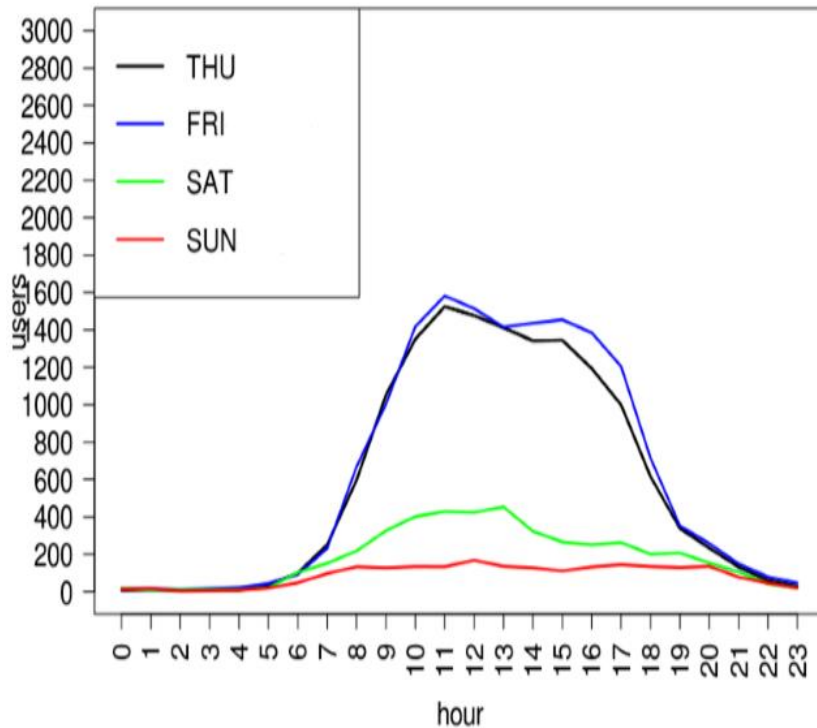
Type X: ?
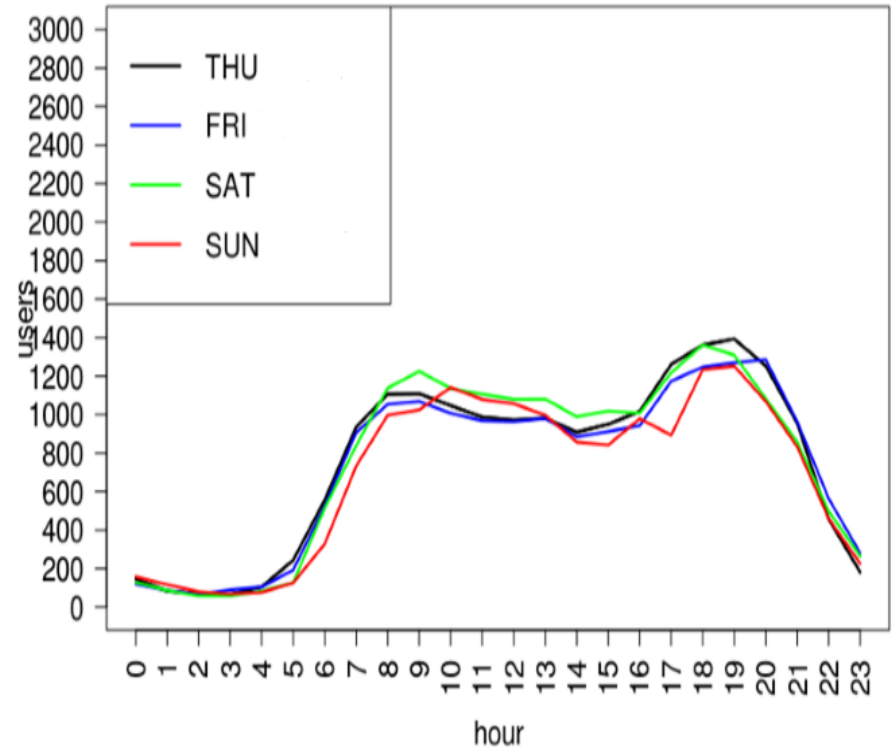
Type Y: ?

- We can use this insight to group base stations into different groups, using unsupervised machine learning techniques

Methodology

- The time series of users connected at a base station contains variations, that can be grouped by similar characteristics
- A month of data is collapsed into an indicative week (Sunday to Saturday), with the time series normalized by the z-score
- Principal Component Analysis(PCA) is used to identify the discriminant patterns from noisy time series data
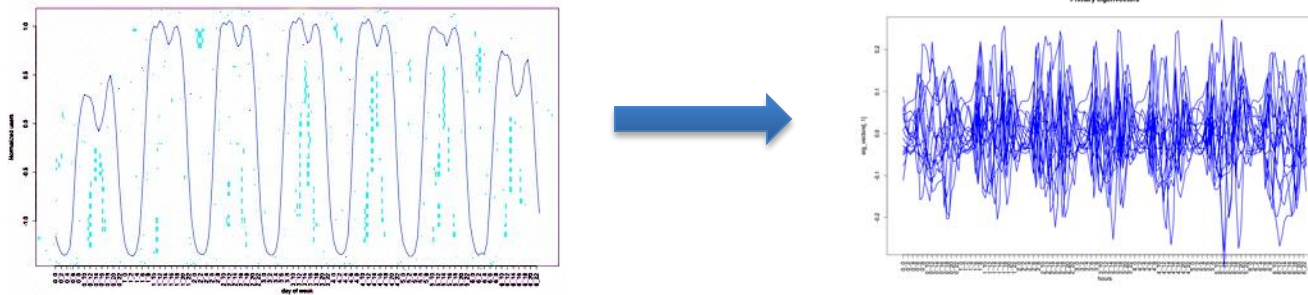
- Each base station's pattern is filtered into 15 principal components (covering 95% of the data for that base station)
- Using the 15 principal components, we cluster all the base stations into 3 clusters in an unsupervised manner using k-means algorithm

Three spatial clusters identified in Colombo District
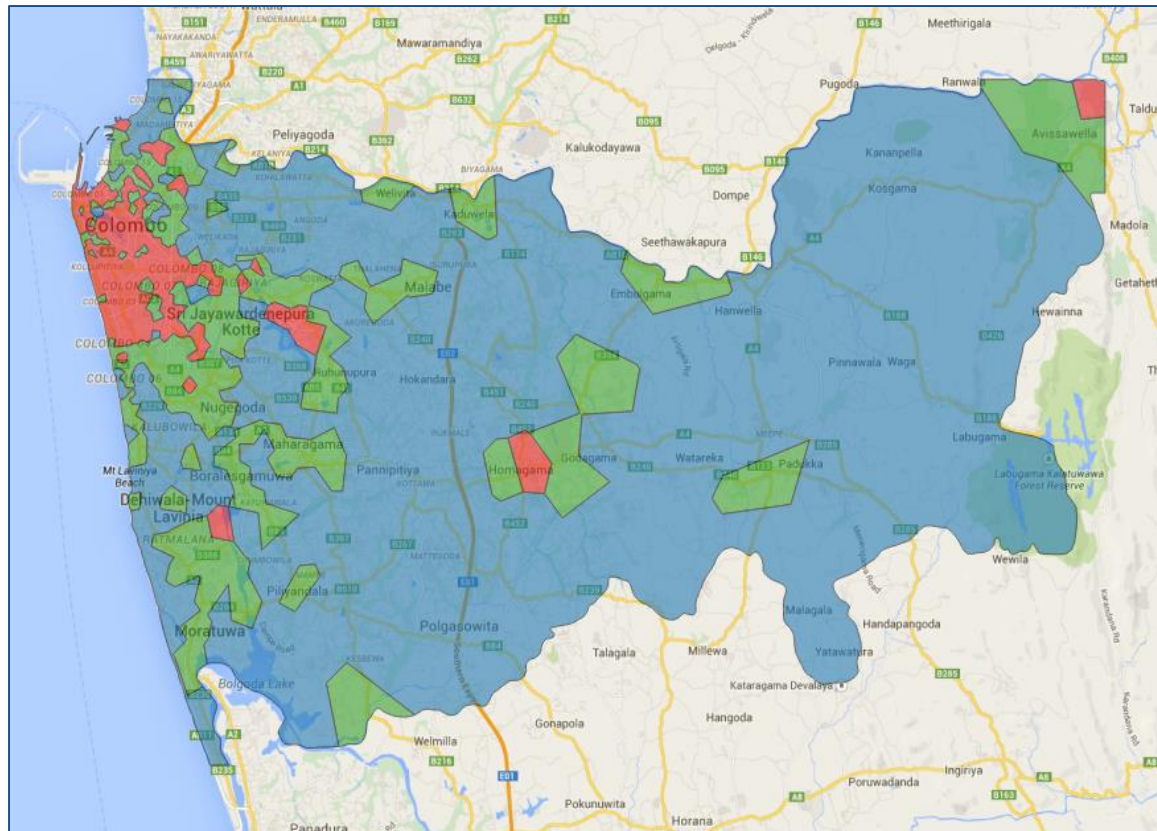
- **Cluster-1 exhibits patterns consistent with commercial area**

- **Cluster-3 exhibits patterns consistent with residential area**

- **Cluster-2 exhibits patterns more consistent with mixed-use**

Our results show Central Business District (CBD) in Colombo city has expanded
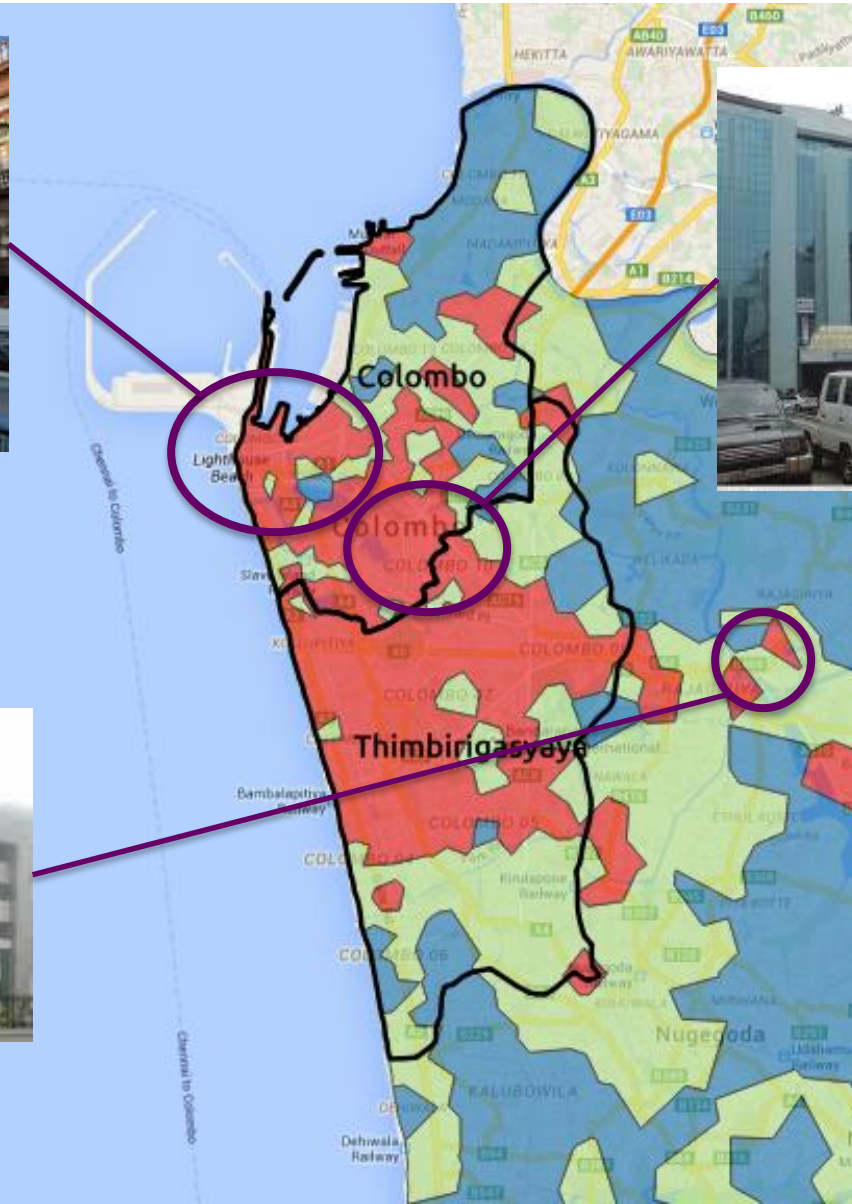
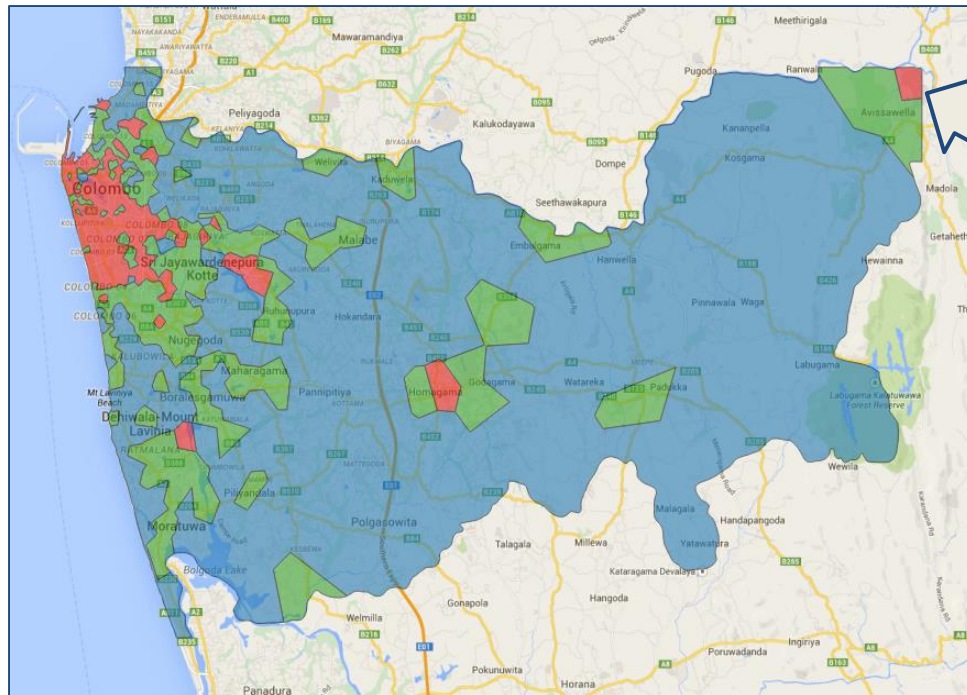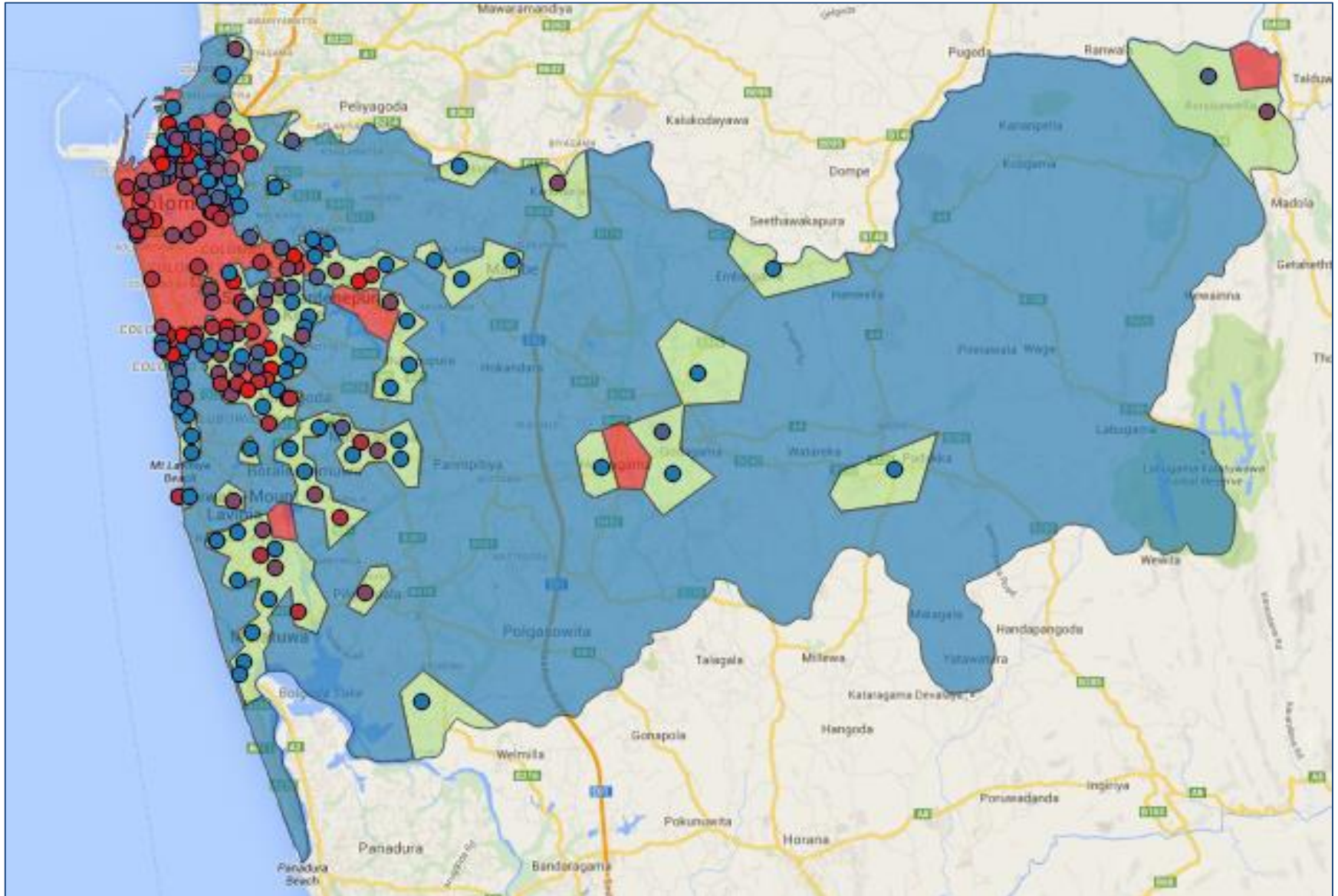# Small area in NE corner of Colombo District classified as belonging to Cluster 1?

Photo ©Senanayaka Bandara - Panoramio

Seethawaka Export Processing Zone

Internal variations in mixed use regions: More commercial or more residential?

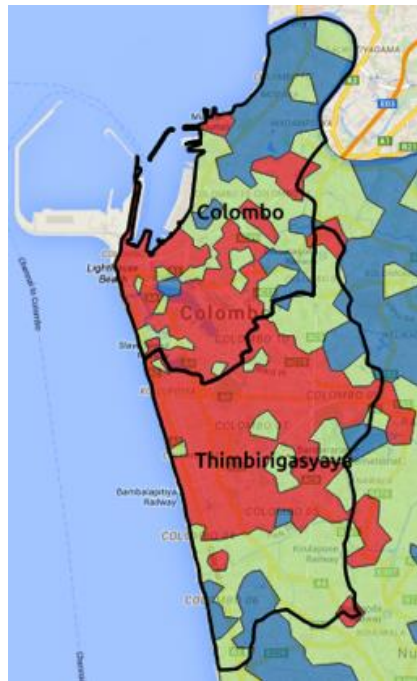Blue dots: more residential than commercial   Red dots: more commercial than residential

Calibrating MNBD based findings with land-use survey data can greatly improve results

- Supervised learning methods can be used to investigate the relationship between zonal land use data and mobile activity data
  - A model trained on areas which have land use data at a higher spatial resolution can be used to predict the land use of other regions
  - Incorrectly labeled cells can reveal how their zonal land use category differ from how they are being used
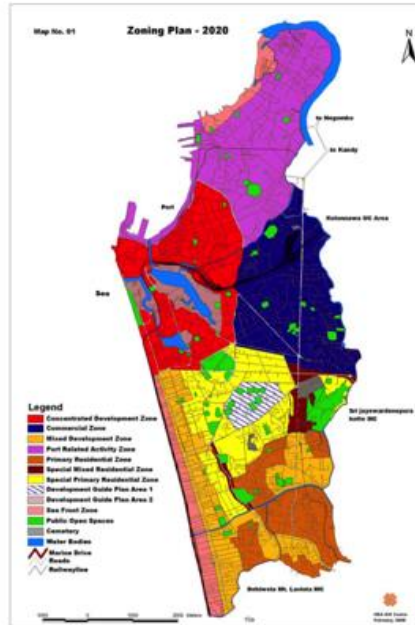
Available "ground-truth" data sources

- Land-use maps compiled by the Survey Department
  - Last updated in 2010; 2014/15 update not yet public
  - Lack of detail on urban land use categories:
    - Urban commercial areas all come under the category of "built-up areas"
    - Urban, semi-urban residential areas all grouped under the category "home gardens"
- Land-use maps compiled by Urban Development Authority
  - Last updated in 2000; 2005 update for Dehiwala-Mt Lavinia MC; 2014/15 update not yet complete
  - Has more relevant classifications for urban areas, but still does not mesh with planned zoning maps

2013 MNBD analyses

2020 UDA plan

2010 Survey Dept. Land Use Map

2000 UDA land use survey

How are we proceeding now?

- Working with Urban Development Authority to calibrate MNBD-based results

- Collaborating with researchers from University of Moratuwa's Town & Country Planning Department to build improved predictive models for spatial changes using MNBD, UDA data, & Survey Dept. data

- Exploring the use of satellite imagery & location based social network data (e.g., Foursquare)
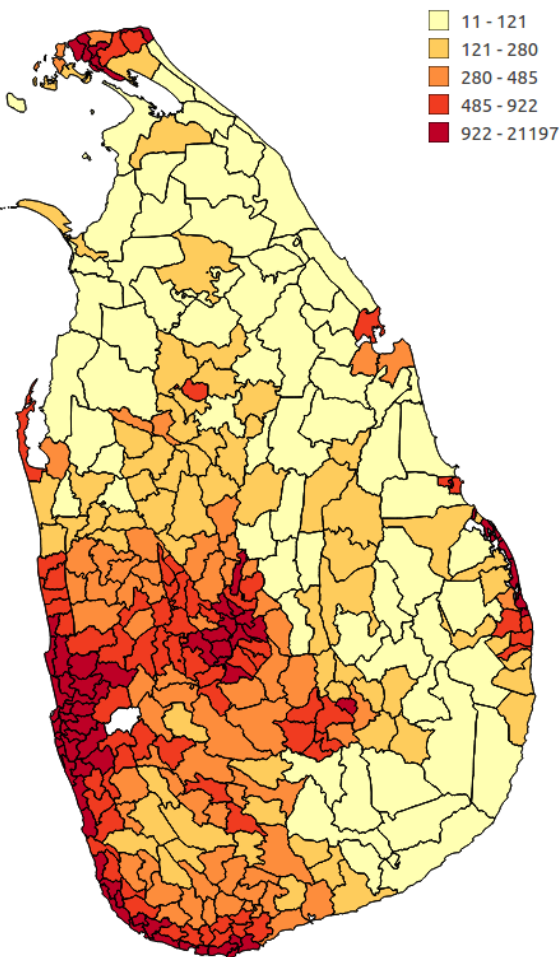 - UoM student project mentored by LIRNEasia

Implications for urban policy

- Almost real-time monitoring of urban land use
- Can complement infrequent surveys & align master plan to reality

EXAMPLE: POPULATION DENSITY & MOVEMENT

MNBD data can give us granular & high-frequency estimates of population density

DSD population density from 2012 census

☐	11 - 121
☐	121 - 280
☐	280 - 485
☐	485 - 922
☐	922 - 21197

DSD population density estimate from MNBD

☐	24 - 141
☐	141 - 272
☐	272 - 479
☐	479 - 1094
☐	1094 - 15088
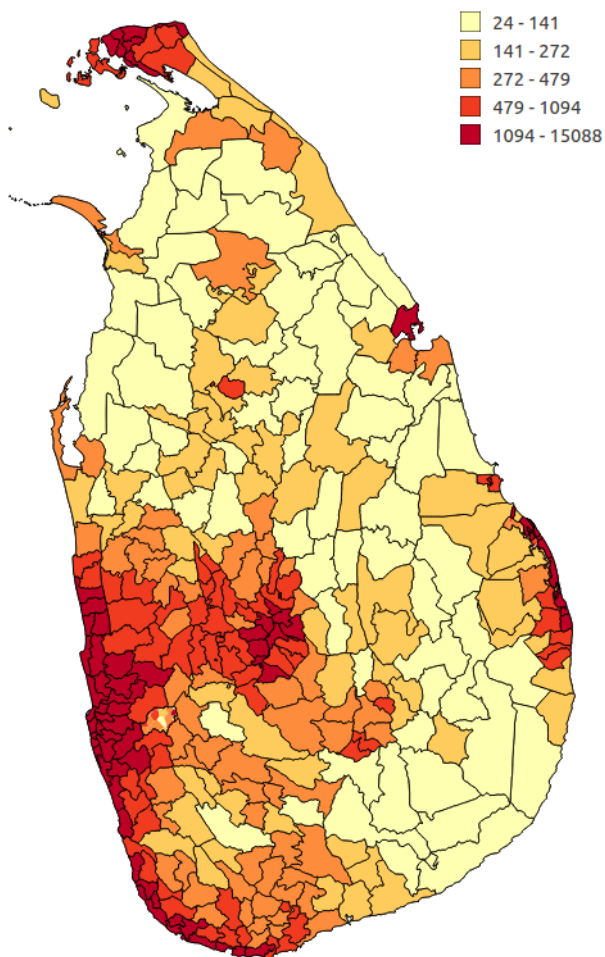
Voronoi cell population density estimate from MNBD

☐	0 - 31
☐	31 - 105
☐	105 - 314
☐	314 - 1134
☐	1134 - 25075
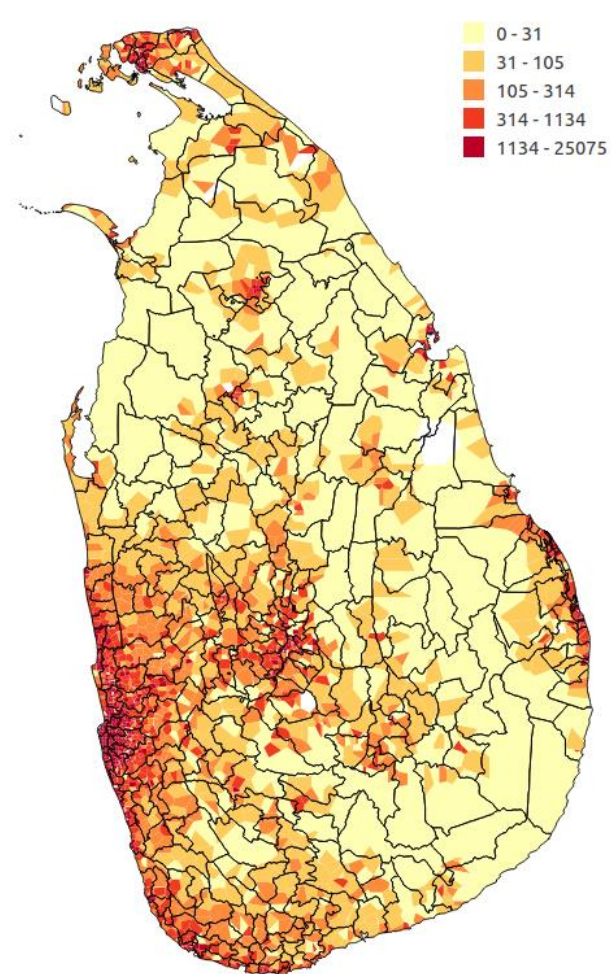
Population density changes in Colombo region: weekday/ weekend

Pictures depict the change in population density at a particular time relative to midnight

Weekday
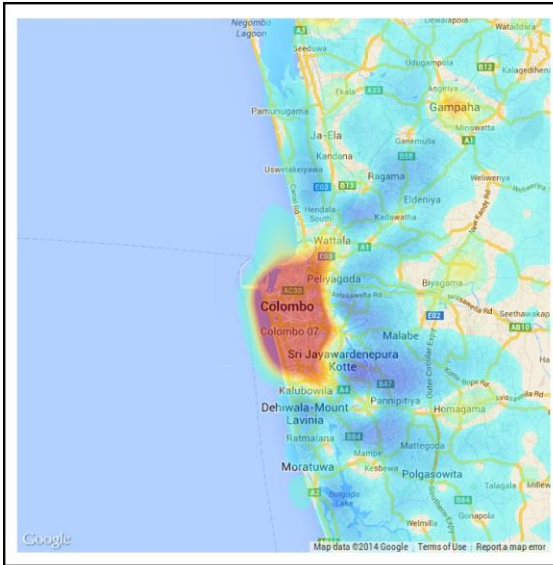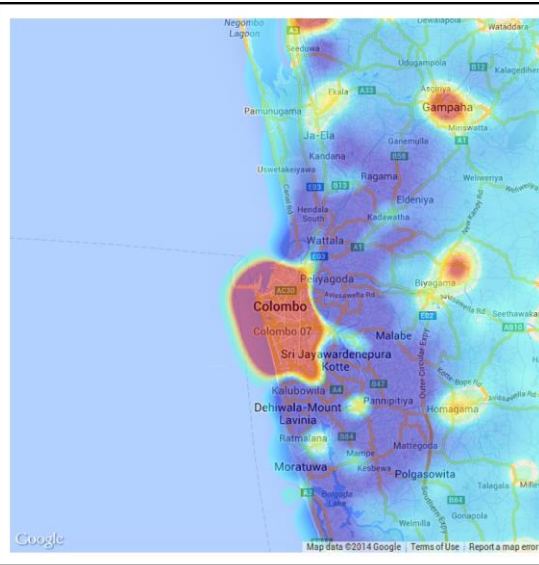
Sunday

Time 06:30   Time 12:30   Time 18:30

Decrease in Density   Increase in Density

24
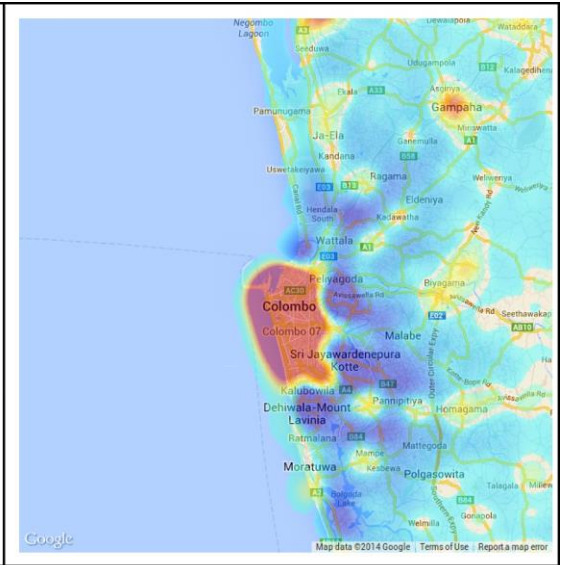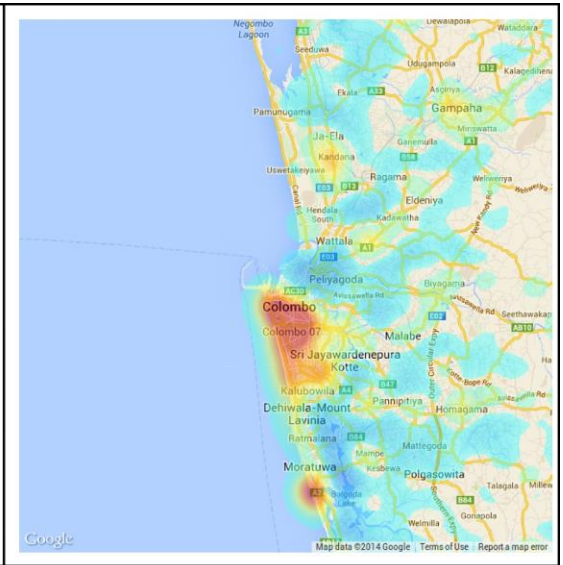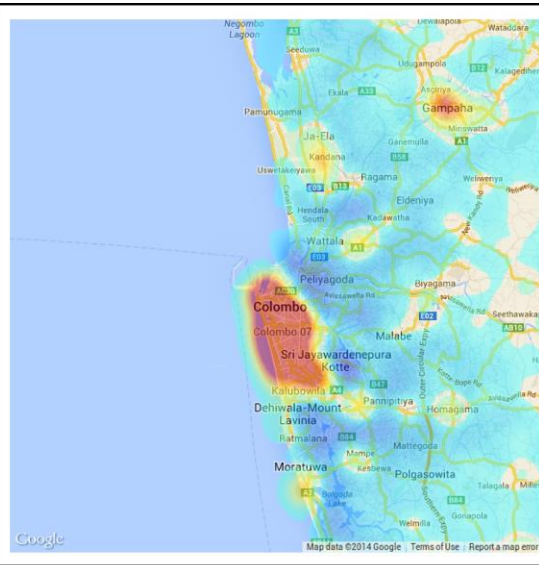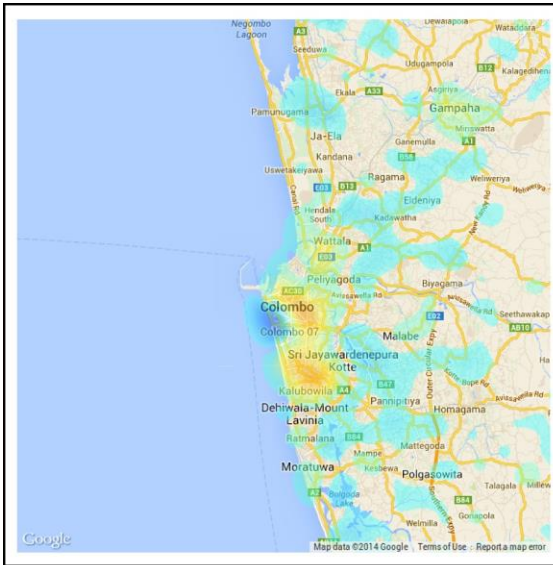
Population density changes in Jaffna & Kandy regions on a normal weekday

Pictures depict the change in population density at a particular time relative to midnight

Jaffna

| Time 06:30 | Time 12:30 | Time 18:30 |

Kandy
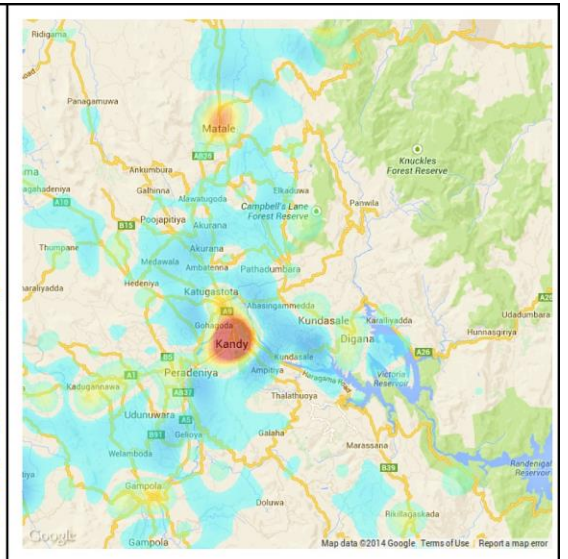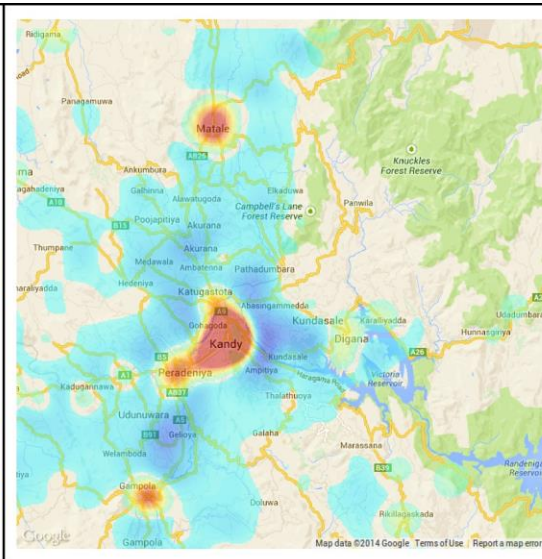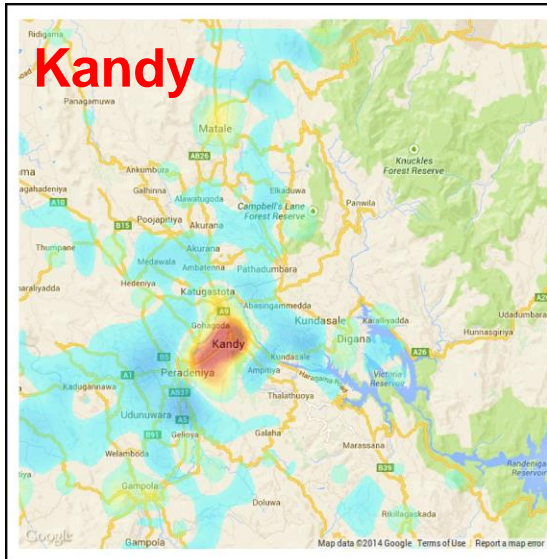
Decrease in Density   Increase in Density

Our findings closely match results from expensive & infrequent transportation surveys; are cheaper & can be produced as needed

26
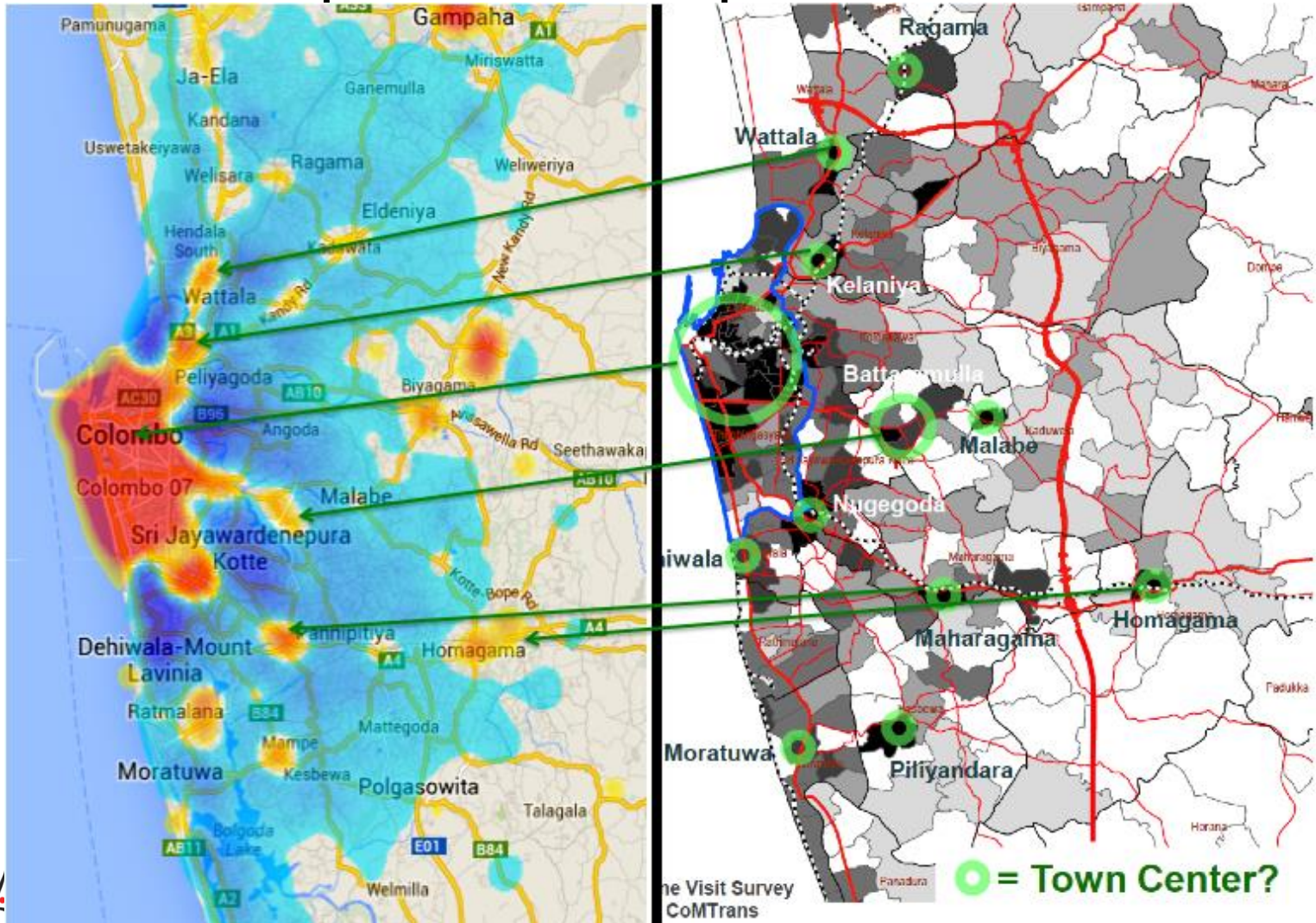
Methodology

- Based on extracted average diurnal mobility pattern for population, choose time frames for home and work
  - Home time: 21:00 to 05:00
  - Work time: 10:00 to 15:00
- Calculate a home and work location for each user:
  - Match cell towers to Divisional Secretariat Division (DSD)
  - Count each DSD at most once per *day*.
  - Pick the DSD with the largest number of "hits"
  - For work consider only weekdays that are not public holidays

46.9% of **Colombo City's** daytime population comes from surrounding DSDs, but some surprises . . .
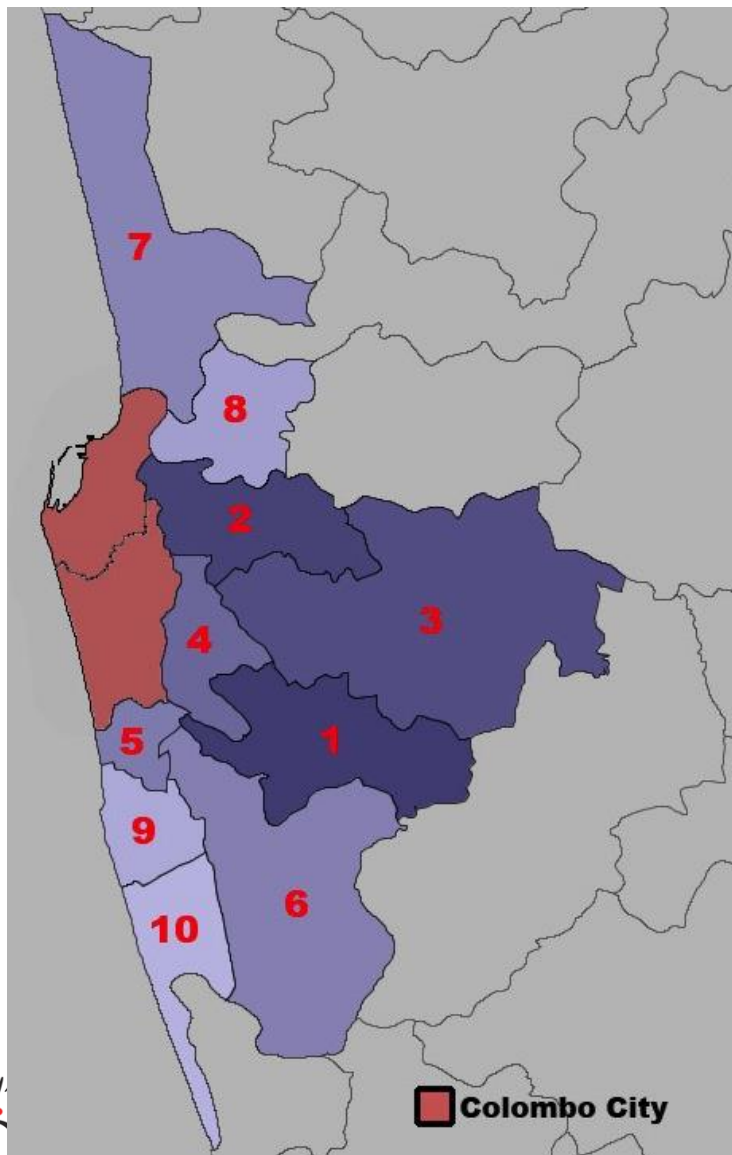
Colombo city is made up of Colombo and Thimbirigasyaya DSDs

Home DSD	%age of Colombo's daytime population
Colombo city	53.1
1. Maharagama	3.7
2. Kolonnawa	3.5
3. Kaduwela	3.3
4. Sri Jayawardanapura Kotte	2.9
5. Dehiwala	2.6
6. Kesbewa	2.5
7. Wattala	2.5
8. Kelaniya	2.1
9. Ratmalana	2.0
10. Moratuwa	1.8

Colombo City

Do current local-government boundaries make sense?

- Commuters cause costs for the city which are not reflected in city revenues

- Fragmented governance makes coherent planning for transport, flood control, solid waste management, etc. difficult

LIRNE*asia*
Pro-poor. Pro-market.

Potential configurations of Metropolitan Colombo Corporation
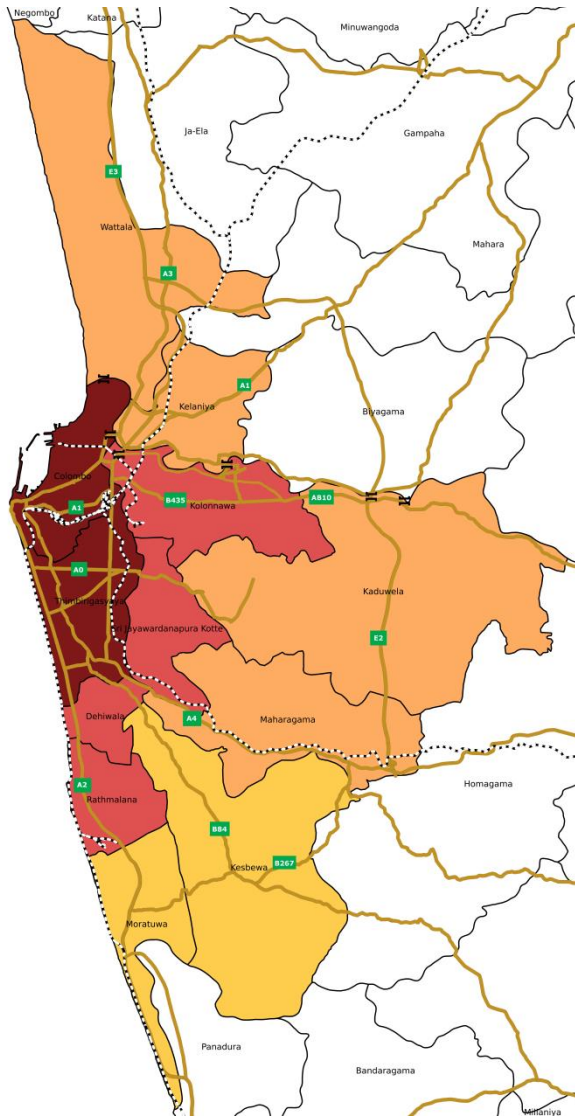
Home DSD	Population	Percentage contribution to Colombo's daytime population	Percentage of home DSD population in located in Colombo during daytime
Colombo City (2 DSDs)	555,031	53.1	86.6
Maharagama	195,355	3.7	17.7
Kolonnawa	190,817	3.5	23.6
Kaduwela	252,057	3.3	17.0
Sri J'pura Kotte	107,508	2.9	24.3
Dehiwala	187,834	2.6	22.9
Kesbewa	244,062	2.5	14.5
Wattala	174,336	2.5	16.1
Kelaniya	134,693	2.1	19.7
Rathmalana	95,603	1.9	18.6
Moratuwa	167,160	1.8	14.7
Total	**2,204,015**	**79.9**	

CHALLENGE 1: POLICY IMPACT

From supply-push to demand-pull

	Event
2014 Oct	• Founding Chair has one-on-one meeting with Secretary, Urban Development ➜ presentation to urban development professionals in November 2014 agreed upon, but postponed due to early announcement of Presidential Election
2015 Jan	• Big Data team conducts a public lecture organized by Institute of Engineers Sri Lanka (IESL)
2015 Feb	• Email contact made with new DG of UDA with offer to brief on LA's ongoing research (meetings planned but don't happen) • First media interactions in Sri Lanka
2015 May	• Big Data Team Leader makes 5-minute presentation at Workshop on implementation of transportation master plan for Ministry of Internal Transport organized by UoM's Dept. of Transport & Logistics. Attended by domain specialists, academics, researchers, officials from UDA, RDA, etc.
2015 Jun - Sept	• Big Data Team Leader invited to join planning team for the Western Region Megapolis Planning Project to provide insights from MNBD. Our name suggested by one of the attendees in the May 2015 event, who was appointed to lead one of the committees working on the WRMPP • In total 9 meetings were attended, culminating in a presentation to all the committees on insights from LIRNEasia big data research related to urban and transportation planning
2015 Aug	• Big Data Team Leader presents at Workshop on Integrated Land Use Transport Modeling Practices in Sri Lanka and around the World organized by UoM's Transportation Engineering Division. Attended by domain specialists, academics, researchers, officials from UDA, RDA, etc.

Policy Enlightenment / Supply-push

Demand-Pull

32

Sri Lanka's highest-circulation English newspaper

Sunday Times 2

Sunday April, 12, 2015

Big data can make South Asian cities smarter

Big data analysis can help citizens make smart choices and plan cities, says Nalaka Gunawardene

View(s):

With more than half of humanity living in cities, there is an urgent need to improve urban planning, design and management.

Until now, policy makers and planners have struggled to keep up with trends. Changes were too fast, and variables too dynamic, for conventional surveys or censuses to capture them adequately.

From supply-push to demand-pull

	Event
2015 Sept	• Big Data Team Leader attends launch of World Bank Report on "Leveraging Urbanization in South Asia" • Side discussions with DG of UDA on LIRNEasia's ongoing research leads to one-on-one meeting with DG the following day to brief him on LIRNEasia research
2015 Dec	• UDA, UDA's Professionals Association, & Young Planners Forum of Institute of Town Planners, Sri Lanka organize special session for LA to present ongoing research. • Echelon Magazine report on Megapolis plans include charts given by UDA on source locations of Colombo's daytime population developed by LIRNEasia (without acknowledgement)
2016 Jan	• Big Data team invited to make presentation to Sri Lanka Strategic Cities Development Project working on Kandy
2016 Feb	• DG UDA requests additional mobility and land-use insights on Kandy
2016 Feb	• Sri Lanka Strategic Cities Development Project reaches out to LIRNEasia for insights on foot traffic in Kandy. Our data not suitable but we brainstorm possible methodologies
2016 Aug	• UDA requests additional finer-grained mobility insights for specific areas in Western Province

Policy Enlightenment / Supply-push

Demand-Pull

LIRNEasia
Pro-poor. Pro-market.
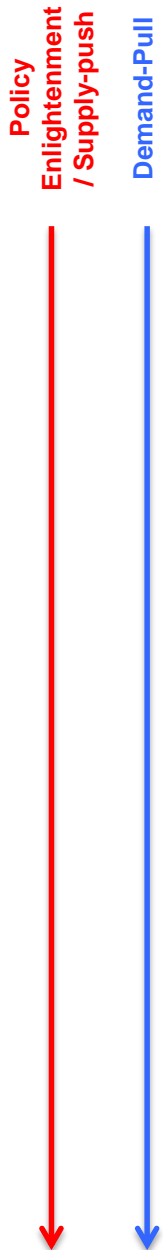
Western Regional Megapolis Project (WRMP)

CROWDED COLOMBO:

Nearly **30%** of the day time population of the city of Colombo, the country's commercial capital, comes from surrounding urban areas

PERCENTAGE OF COLOMBO'S DAY TIME POPULATION FROM

1. Maharagama	**4%**	6. Kesbewa	**3%**
2. Kolonnawa	**4%**	7. Wattala	**3%**
3. Kaduwela	**3%**	8. Kelaniya	**2%**
4. Kotte	**3%**	9. Ratmalana	**2%**
5. Dehiwala	**3%**	10. Moratuwa	**2%**

CONNECTING PEOPLE:

Transport corridors have been identified connecting residential areas with the city centre in Colombo

GROWING POPULATION:

The population of the Western Province comprising Colombo, Gam and Kalutara districts was nearly 4 million in 1981 and will more th **double by 2030**.

AREA SHOWN ON THE MAP

ter of cities. Singaporean government city planners

Source: Interview with Western Region Megapolis Authority in Echelon magazine (December 2015, pp. 63)

Key lessons

- No demand for insights from big-data research when we started in 2012; but had we not started on the research when we did, no results would have been available when the policy window opened

- Given difficulty of assessing quality of big-data research, the credibility of LIRNEasia has been of value
  - More work needs to be done to make potential users of this research more knowledgeable

- Supply-push approach helped create the conditions for essential demand-pull
  - However, political changes and appointments which were outside our control were critical in creating demand

LIRNE*asia*
Pro-poor. Pro-market.

CHALLENGE 2: HUMAN RESOURCES

Data scientists are in short supply

- Workaround: Multi-disciplinary teams
  - We prioritize analytical thinking over knowledge of big data tools in our recruitment interviews
  - Team members have different specialties
    - Staff and consultants have a mix of computer-science skills, statistics, and domain knowledge

Staff

- Danaja Maldeniya (Computer Science, Statistics)
  - Now at U of Michigan but still collaborating
- CD Athuraliya (Computer Science)
- Dedunu Dhananjaya (Software Engineer, Systems Administrator)
  - moved to private sector, but still collaborating
- Isuru Jayasooriya (Computer Science, Machine Vision)
- Kaushalya Madhawa (Computer Science, Statistics)
  - Now at Tokyo Institute of Technology
- Keshan de Silva (Computer Science, Agent Based Simulations)
- Lasantha Fernando (Computer Science)
- Madhushi Bandara (Computer Science)
  - Now at U of New South Wales but still collaborating
- Nisansa de Silva (Computer Science)
  - Now at U of Oregon, but still collaborating
- Robert Galyean (Mathematics, Physics)
- Prof. Rohan Samarajiva (Public Policy)
- Sriganesh Lokanathan (Public Policy, Computer Science)
- Shazna Zuhyle (Research Manager)
- Thavisha Gomez (Research Manager)

Collaborators

- Prof. Amal Kumarage (Dept. of Transport & Logistics, UoM)
  - Transportation, Urban Planning
- Dr Amal Shehan Perera (Dept. of CSE, UoM)
  - Data Mining
- Fields of View
  - Indian research institute specializing in games and simulations for public policy issues
- Gabriel Kreindler (MIT)
  - Economics
- Prof Joshua Blumenstock (UC Berkley, School of Information)
  - Data Science, Economics
- Prof Moinul Zaber (U of Dhaka)
  - Data Science, Public Policy
- Shibasaki & Sekimoto Laboratory, University of Tokyo
  - Big Data for Development research practice
- Yuhei Miyauchi (MIT)
  - Economics

Recruitment remains a challenge

- A fresh computer science undergrad does not think of working in a think-tank; often looking to join a software firm
  - LIRNE*asia* has done presentations in public forums and at universities to broaden horizons
- In people's minds "research" often means low compensation
  - Our packages are competitive with private sector, but ultimately we are a not-for-profit
- Key selling point is the work that we do and our partnerships
  - It is rewarding to see research being used
  - Good opportunities for publication and conference papers
  - LIRNEasia encourages individual researchers to build their "brands"
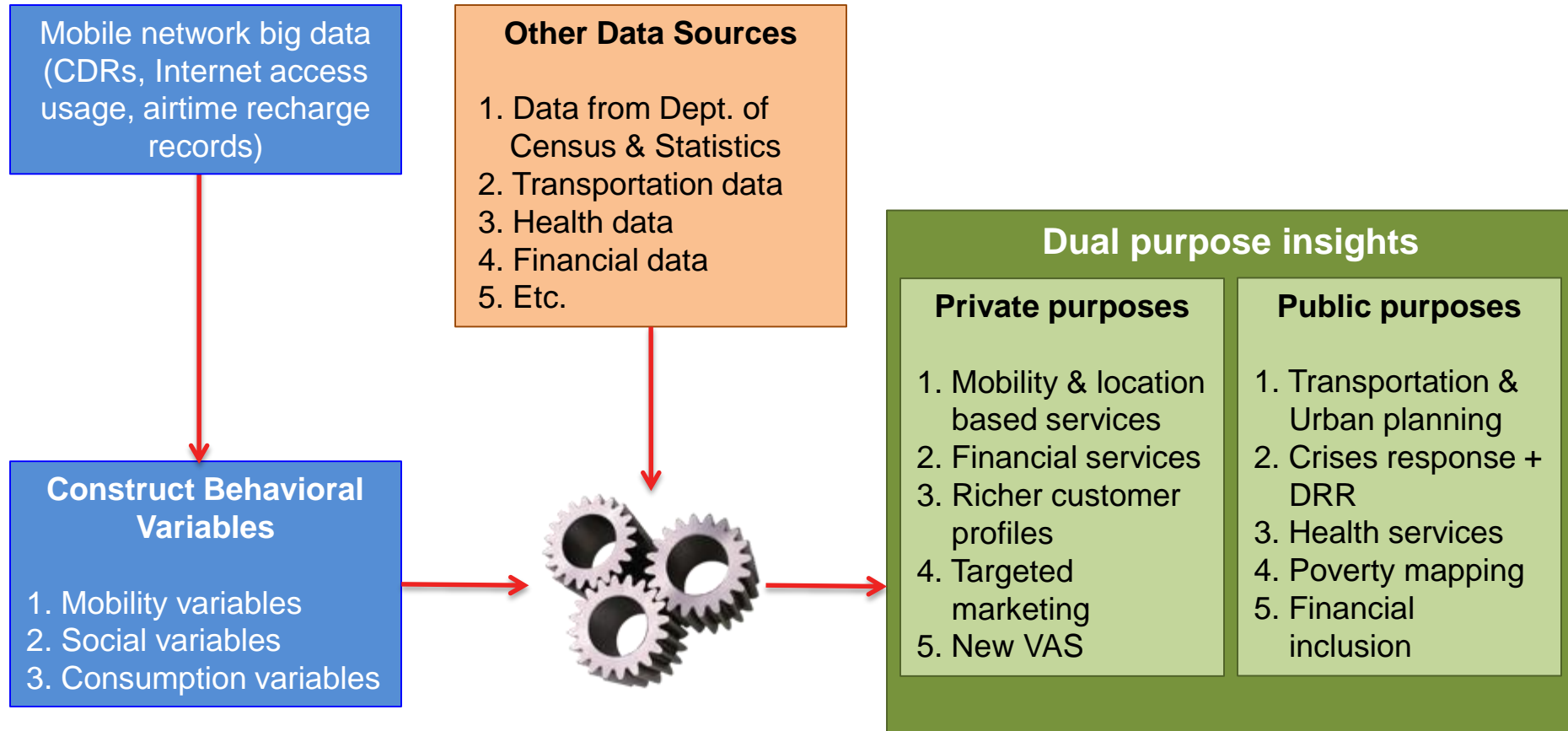 - Ideal for people looking to get into good PhD programs

LIRNE*asia* Data Analytics Advisory Council

- **Dr. Linus Bengtsson, PhD**

  Co-Founder and Executive Director, Flowminder

- **Joshua Blumenstock, PhD**

  Assistant Professor, School of Information

  Director, Data Science and Analytics Laboratory

  University of Washington

- **Nitesh Chawla, PhD**

  Professor of Computer Science and Engineering

  Director of The Interdisciplinary Center for Network Science & Applications (iCeNSA)

  University of Notre Dame

- **Vanessa Frias-Martinez, PhD**

  Assistant Professor, School of Information Studies

  University of Maryland

- **Amal Kumarage, PhD**

  Senior Professor at Department of Transport & Logistics Management, University of Moratuwa

- **Ashwin Mahesh, PhD**

  Founder & CEO, Mapunity

- **P.K.S. Mahanama, PhD**

  Professor and Former Dean of Department of Town and Country Planning, University of Moratuwa

- **Wasan Pattara-atikom, PhD**

  Principal Researcher & Head of Intelligent Transportation System Laboratory, NECTEC

- **Amal Shehan Perera, PhD**

  Senior Lecturer, Department of Computer Science and Engineering, University of Moratuwa

- **Louiqa Raschid, PhD**

  Professor, Smith School of Business, Center for Bioinformatics and Computational Biology

  UMIACS and the Department of Computer Science

  Robert H. Smith School of Business, University of Maryland

- **Srinath Perera, PhD**

  Vice President, Research, WSO2 Inc.

- **Prabir Sen, PhD**

  Former Chief Data Scientist

  Infocomm Development Authority of Singapore

- **Hetan Shah**

  Executive Director, Royal Statistical Society

- **Ryosuke Shibasaki, PhD**

  Professor, Dr.Eng. Center for Spatial Information Science

  University of Tokyo

- **Linnet Taylor, PhD**

  Marie Curie Research Fellow, Faculty of Social and Behavioural Sciences, University of Amsterdam

- **Ruvan Weerasinghe, PhD**

  Senior Lecturer, University of Colombo - School of Computing

- **Arj Wignaraja**

  Vice President of Operations, Remote Sensing Metrics, LLC

CHALLENGE 3: DATA

Mobile network big data + other data ➡ rich, timely insights that serve private **as well** as public purposes

Mobile network big data (CDRs, Internet access usage, airtime recharge records)

Other Data Sources

1. Data from Dept. of Census & Statistics
2. Transportation data
3. Health data
4. Financial data
5. Etc.

Construct Behavioral Variables

1. Mobility variables
2. Social variables
3. Consumption variables

Dual purpose insights

Private purposes	Public purposes
1. Mobility & location based services	1. Transportation & Urban planning
2. Financial services	2. Crises response + DRR
3. Richer customer profiles	3. Health services
4. Targeted marketing	4. Poverty mapping
5. New VAS	5. Financial inclusion

M... k big data + other data ➔ rich, timely insights
... rve private **as well** as public purposes

Private sector data

Government data

Mobile network big data (CDRs, Internet access usage, airtime recharge records)

Other Data Sources

1. Data from Dept. of Census & Statistics
2. Transportation data
3. Health data
4. Financial data
5. Etc.

Construct Behavioral Variables

1. Mobility variables
2. Social variables
3. Consumption variables

Dual purpose insights

Private purposes

1. Mobility & location based services
2. Financial services
3. Richer customer profiles
4. Targeted marketing
5. New VAS

Public purposes

1. Transportation & Urban planning
2. Crises response + DRR
3. Health services
4. Poverty mapping
5. Financial inclusion

44

Obtaining mobile network data

- No established process exists, therefore prior relationships matter

- Basic process

  1. Obtain in-principle agreement from CEOs of companies, ideally at least two

  2. Negotiate specifics with 2nd and 3rd tier management

  3. Approach other operators

- Throughout the process, reciprocity was emphasized

  - Methods for deriving public policy insights can also be adapted for commercial purposes

LIRNEasia
Pro-poor. Pro-market.

Answering operator concerns

- Question: Will the regulator raise any concerns?

- Answer:
    - The specific data requested did not contravene existing laws or license conditions
    - Data we obtain is pseudonymized with no links to original numbers

- Question: Will this research reveal any proprietary business intelligence?

- Answer:
    - We combine data from multiple operators
    - All researchers sign NDAs with LIRNEasia
    - Operators sign off on our results before public release

LIRNEasia
Pro-poor. Pro-market.

Establishing a data-sharing agreement with an operator

- One short initial meeting between LIRNEasia's Founding Chair and CEO
  - Had sent a brief concept note on potential collaboration prior to meeting
  - Obtained in-principle agreement, followed by letter
- Negotiate specifics with 2nd and 3rd tier management:
  - 7 in-person meetings
    - Marketing, Business Intelligence, Network Engineering, Regulatory & Legal
    - ~7.5 hours in total
  - 4 conference calls
    - ~1 hour in total with operator; ~1 hour in total with LIRNEasia lawyers
  - 16 email exchanges
    - 11 with operator; 5 with LIRNEasia lawyers
  - 6 rounds of revisions of basic agreement
  - 11 people involved in total
    - 7 from operator; 3 from LIRNEasia; 1 from LIRNEasia's lawyers
- Sign agreements

6 months

How can access to data for research for public purposes be made easier?

- Reduce transaction costs:
  - Standardized agreement template(s)
- Pro-actively deal with privacy concerns
  - Operators adopt self-regulatory guidelines for minimizing the potential harms from giving access to mobile network data
  - LIRNEasia working with operators in the region in this regard
    - Draft self-regulatory guidelines developed by LIRNEasia available at http://lirneasia.net/2014/08/what-does-big-data-say-about-sri-lanka/
- Perhaps there is a need for new operational models with third-party data guardians
  - E.g., Yale University Open Data Access (YODA) project acting as data guardians and providing researchers access to Johnson & Johnson clinical trial data

Obtaining government data

- Some processes exist for getting data, but uniform procedures do not
  - Easier for organizations which are affiliated with government, which LIRNEasia is not
- Different approaches
  - Submitting a data request proposal
  - Meeting(s) with senior officials
  - Spending time talking to junior officials and negotiating access
  - Building a compelling story from analyses of mobile network data showing how insights can be improved with use of government data
  - Partnering with government-affiliated organization (e.g. UoM)