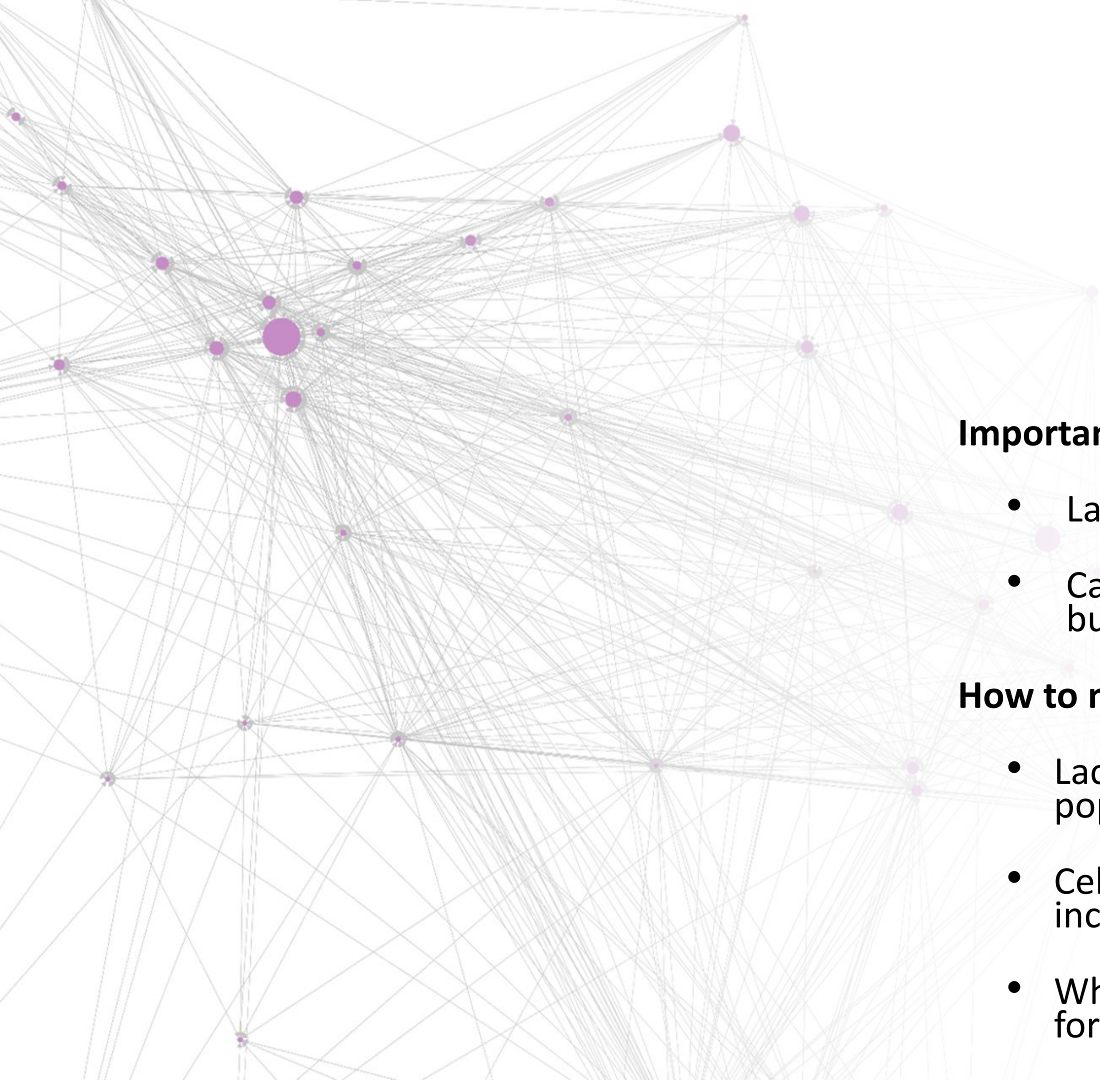
Telecom Data Processing and Analysis

Sveta Milusheva, Leonardo Viotti and Dunstan Matekenya

Outline

- Why focus on telecom data (CDR)?
- What is CDR data?
- Extraction and Aggregation of Data
- Quality Checks
- Analysis and Visualizations
- Practical exercises

CDR and mobility



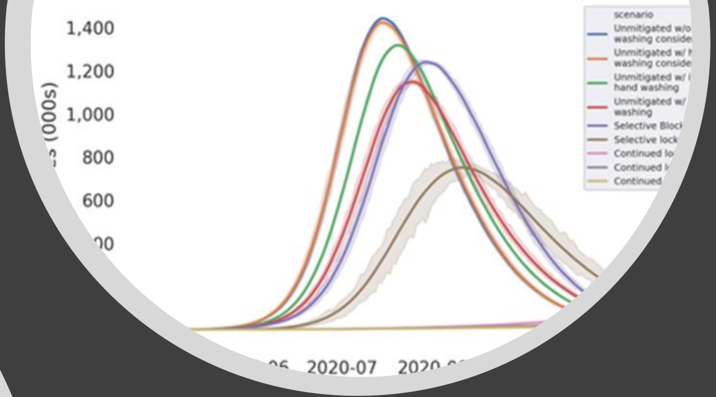
Mobility and Density

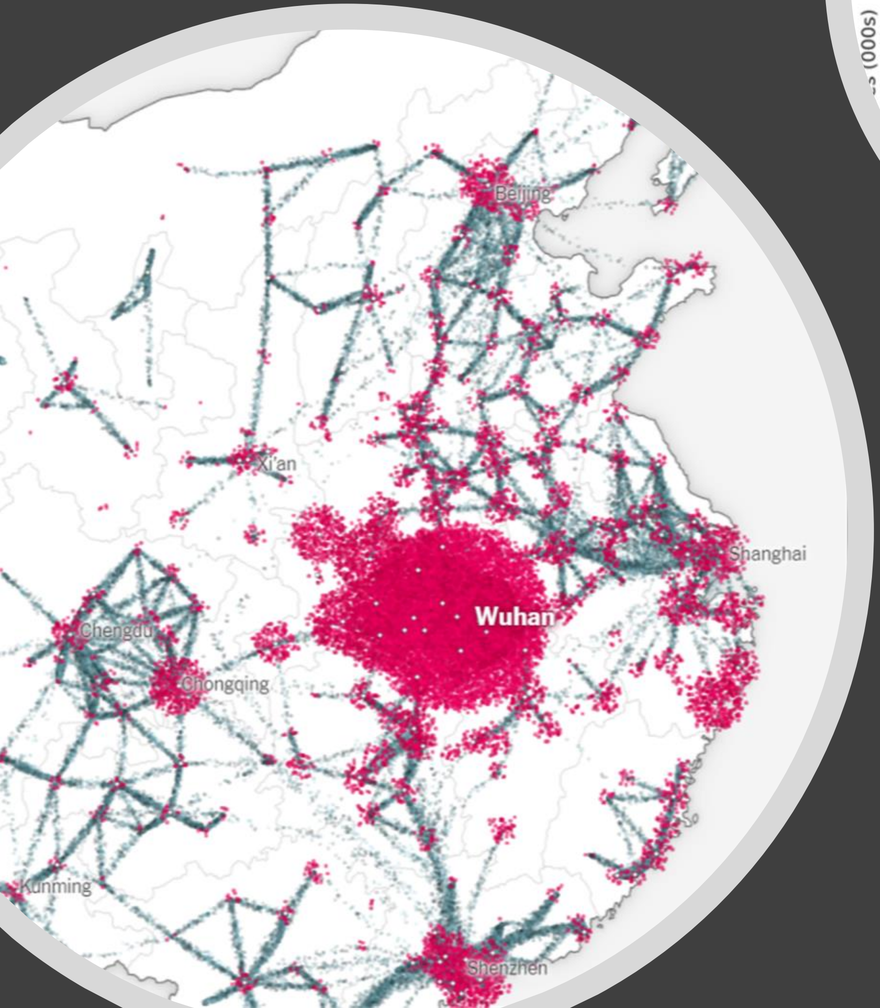
Importance of Mobility

- Large increases in mobility in recent years
- Can lead to spread of information & technology, but also disease

How to measure?

- Lack of information about mobility on a population level
- Cell phones are ubiquitous even among low-income
- Where people use their cellphones is a good proxy for their current location

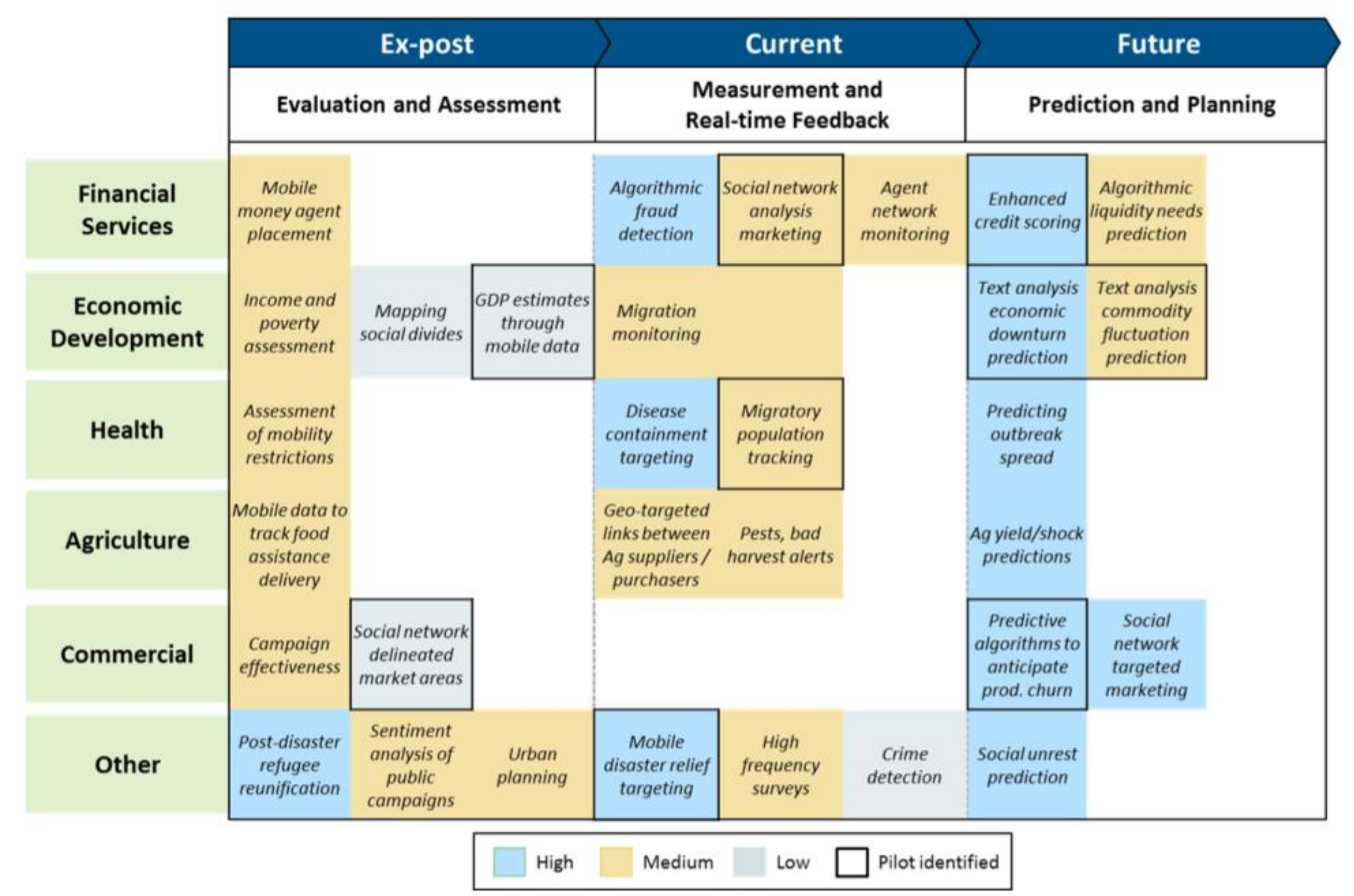




Especially important for COVID19

Use mobility to inform modeling to understand spread of cholera and other diseases

Broad Application Areas



	Ex-post			Current			Future	
	Evaluation and Assessment			Measurement and Real-time Feedback			Prediction and Planning	
Financial Services	🟨 Mobile money agent placement			🟦 Algorithmic fraud detection	🟨 □ Social network analysis marketing	🟨 Agent network monitoring	🟦 □ Enhanced credit scoring	🟨 Algorithmic liquidity needs prediction
Economic Development	🟨 Income and poverty assessment	▨ Mapping social divides	▨ □ GDP estimates through mobile data	🟨 Migration monitoring			🟦 □ Text analysis economic downturn prediction	🟨 □ Text analysis commodity fluctuation prediction
Health	🟨 Assessment of mobility restrictions			🟦 Disease containment targeting	🟨 □ Migratory population tracking		🟦 Predicting outbreak spread	
Agriculture	🟨 Mobile data to track food assistance delivery			🟨 Geo-targeted links between Ag suppliers / purchasers	🟨 Pests, bad harvest alerts		🟦 Ag yield/shock predictions	
Commercial	🟨 Campaign effectiveness	▨ □ Social network delineated market areas					🟦 □ Predictive algorithms to anticipate prod. churn	🟦 Social network targeted marketing
Other	🟦 Post-disaster refugee reunification	🟨 Sentiment analysis of public campaigns	🟨 Urban planning	🟦 □ Mobile disaster relief targeting	🟨 High frequency surveys	▨ Crime detection	🟦 Social unrest prediction	

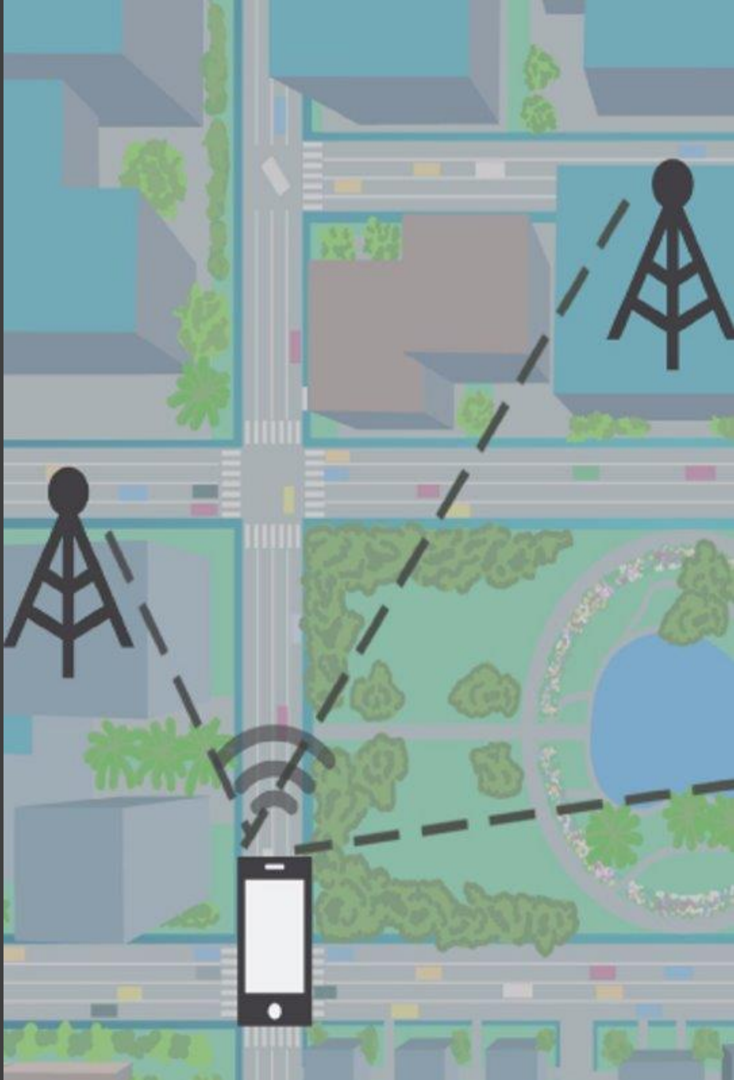
🟦 High 🟨 Medium ▨ Low □ Pilot identified

Source: https://docs.gatesfoundation.org/Documents/Using%20Mobile%20Data%20for%20Development.pdf

What is CDR data?

Cellular Networks

- Mobile phone networks (e.g., GSM, CDMA, LTE) require regular pings between mobile phone devices and cellular communication antennas. The networks constantly determine the location of the mobile phone devices even when the device is on standby.
- Two types of location updates
  - Network triggered updates
    - being switched on and connects to the cellular network
    - involved in a call and moves between two different cell areas (i.e. cell handover)
  - Event triggered updates: based on device usage (e.g., call, SMS, internet use, apps)

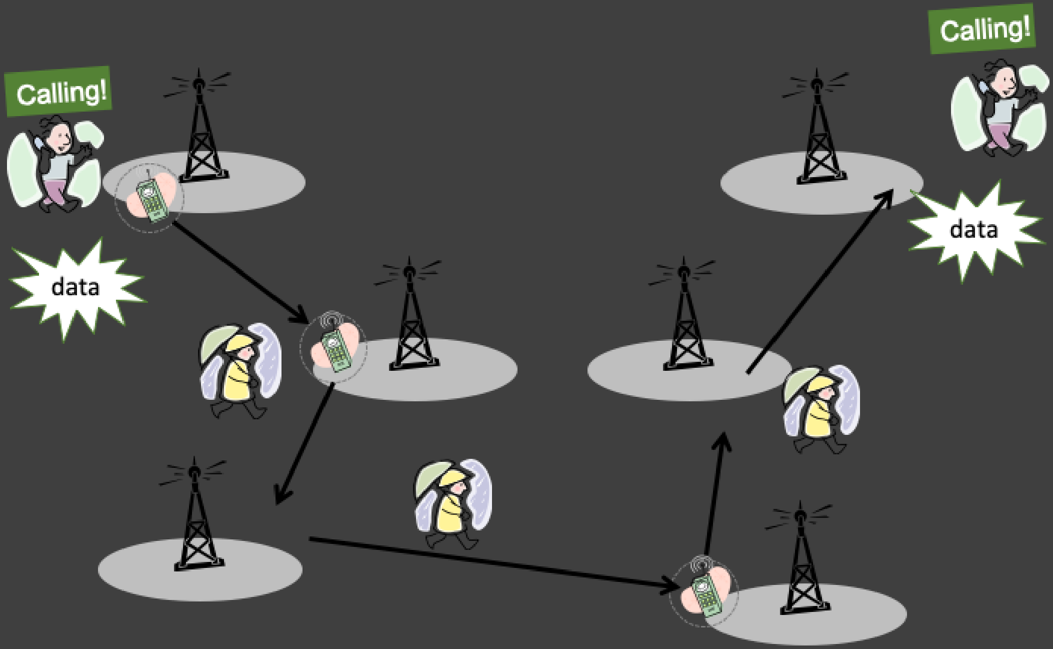


Call Detail Records (CDR)?

Call Detail Records or CDRs are created whenever an individual interacts with the mobile network (event-triggered updates). They are used by MNOs for billing purposes.

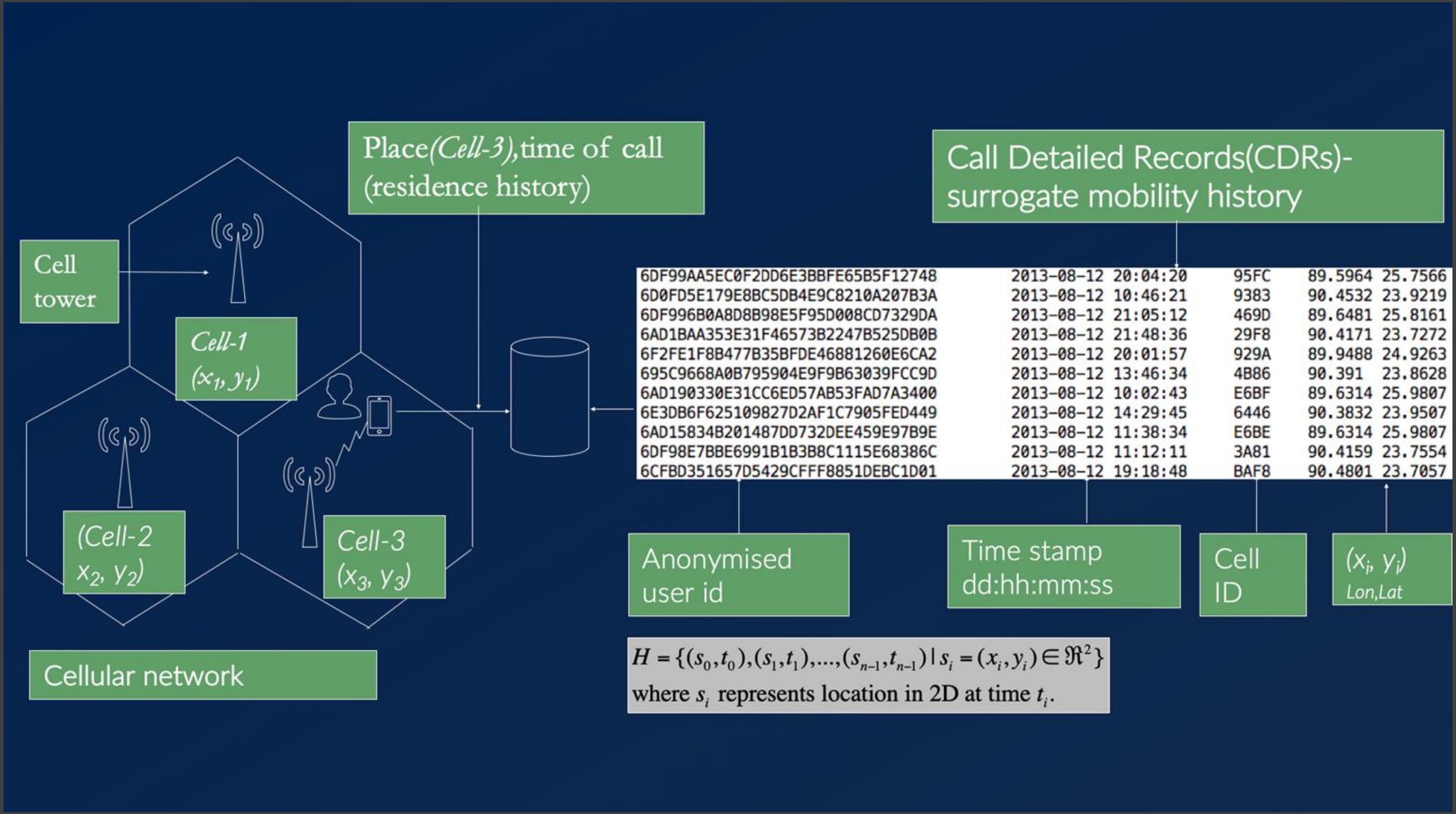
A Typical CDR contains the following:

- Phone numbers (origin, party-A & receiver, party-B)
- Time stamp of the call
- Cell tower (as latitude, longitude) through which the call entered and left the exchange
- Call type (e.g., voice, SMS, internet)
- Call duration



Source: University of Tokyo slides

CDR Data Structure



Cell tower

Place*(Cell-3)*,time of call (residence history)

Cell-1 (x_1, y_1)

(Cell-2 x_2, y_2)

Cell-3 (x_3, y_3)

Cellular network

Call Detailed Records(CDRs)- surrogate mobility history

6DF99AA5EC0F2DD6E3BBFE65B5F12748	2013-08-12 20:04:20	95FC	89.5964	25.7566
6D0FD5E179E8BC5DB4E9C8210A207B3A	2013-08-12 10:46:21	9383	90.4532	23.9219
6DF996B0A8D8B98E5F95D008CD7329DA	2013-08-12 21:05:12	469D	89.6481	25.8161
6AD1BAA353E31F46573B2247B525DB0B	2013-08-12 21:48:36	29F8	90.4171	23.7272
6F2FE1F8B477B35BFDE46881260E6CA2	2013-08-12 20:01:57	929A	89.9488	24.9263
695C9668A0B795904E9F9B63039FCC9D	2013-08-12 13:46:34	4B86	90.391	23.8628
6AD190330E31CC6ED57AB53FAD7A3400	2013-08-12 10:02:43	E6BF	89.6314	25.9807
6E3DB6F625109827D2AF1C7905FED449	2013-08-12 14:29:45	6446	90.3832	23.9507
6AD15834B201487DD732DEE459E97B9E	2013-08-12 11:38:34	E6BE	89.6314	25.9807
6DF98E7BBE6991B1B3B8C1115E68386C	2013-08-12 11:12:11	3A81	90.4159	23.7554
6CFBD351657D5429CFFF8851DEBC1D01	2013-08-12 19:18:48	BAF8	90.4801	23.7057

Anonymised user id

Time stamp dd:hh:mm:ss

Cell ID

(x_i, y_i) *Lon,Lat*

$$H = \{(s_0,t_0),(s_1,t_1),\dots,(s_{n-1},t_{n-1}) \mid s_i = (x_i,y_i) \in \Re^2\}$$

where s_i represents location in 2D at time t_i.

CDR and privacy

Since CDR data contains confidential information for the subscribers and the telecom companies, observing strict practices to keep the data safe and private is crucial:

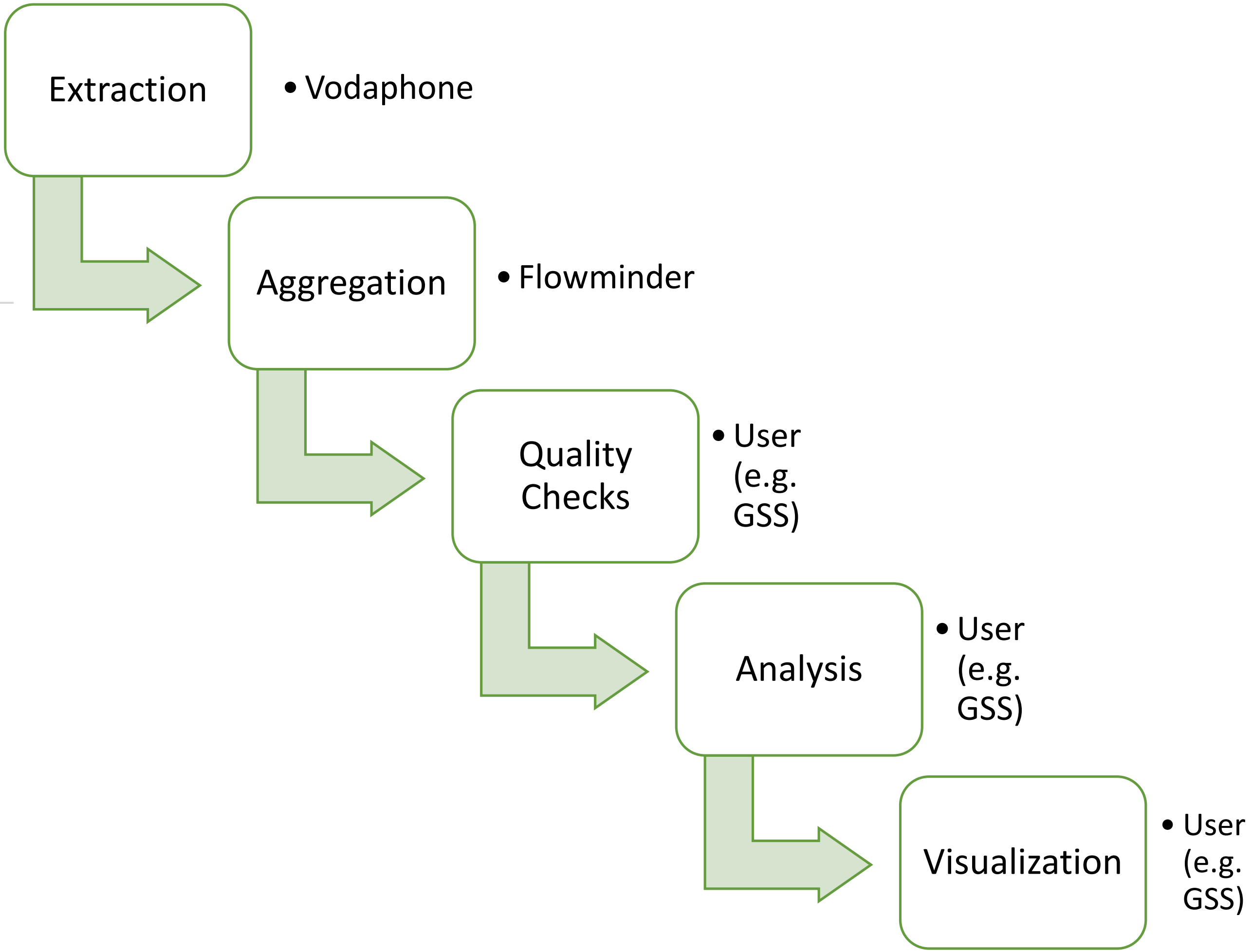
- Often, only anonymized data is analyzed. Anonymization
    - Removes personally identifying information (PII) (e.g., phone numbers and IMEIs)
    - Reduces risk of re-identification of individuals in the data
- Access to individual level data is limited to few people
- Data storage and processing is done within secure environments

CDR limitations

Important biases to consider when interpreting CDR data:

- **Data representativeness** of a population depends on cell phone penetration and usage patterns across different groups
- **Geographic coverage**: Towers more concentrated in urban areas
- **Data from single operator**: Different operators can represent different users and behaviours (e.g. one operator has a more wealthy or urban user-base while another has more users in rural areas)

Overview processing pipeline



Extraction

- Vodaphone

Aggregation

- Flowminder

Quality Checks

- User (e.g. GSS)

Analysis

- User (e.g. GSS)

Visualization

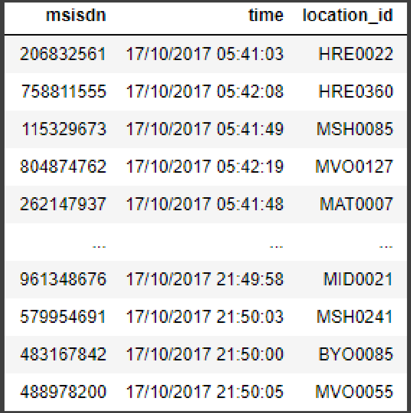
- User (e.g. GSS)

Aggregation

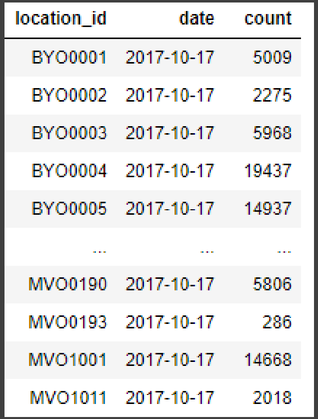
Aggregation and analysis

Aggregations are indicators used for summarizing and interpreting the data.

- One of the most efficient way to anonymize individual level data
- Reduces the size of data, but preserves key information



msisdn	time	location_id
206832561	17/10/2017 05:41:03	HRE0022
758811555	17/10/2017 05:42:08	HRE0360
115329673	17/10/2017 05:41:49	MSH0085
804874762	17/10/2017 05:42:19	MVO0127
262147937	17/10/2017 05:41:48	MAT0007
...	...	...
961348676	17/10/2017 21:49:58	MID0021
579954691	17/10/2017 21:50:03	MSH0241
483167842	17/10/2017 21:50:00	BYO0085
488978200	17/10/2017 21:50:05	MVO0055



location_id	date	count
BYO0001	2017-10-17	5009
BYO0002	2017-10-17	2275
BYO0003	2017-10-17	5968
BYO0004	2017-10-17	19437
BYO0005	2017-10-17	14937
...	...	...
MVO0190	2017-10-17	5806
MVO0193	2017-10-17	286
MVO1001	2017-10-17	14668
MVO1011	2017-10-17	2018

Aggregation and analysis

Here are some common types of aggregation we can create from raw CDR data:

- Number of transactions per day and region
- Number of active subscribers per day and region
- Number of movements between two regions per day

We will see examples of each on the practical exercises.

Data Quality Checks

CDR Data Quality Challenges

Causes

- When towers are deactivated (e.g for maintenance) transactions are redirected to other towers.
- Errors with initial data extraction

How These may impact the Data

- Sudden spikes or drops in a certain region
- Missing data or duplicates
- Much fewer observations at periods in time

Quality checks

Completeness

Check if all expected data is present.

- Does it cover the entire time period?
- Does it include all geographic regions?

Consistency

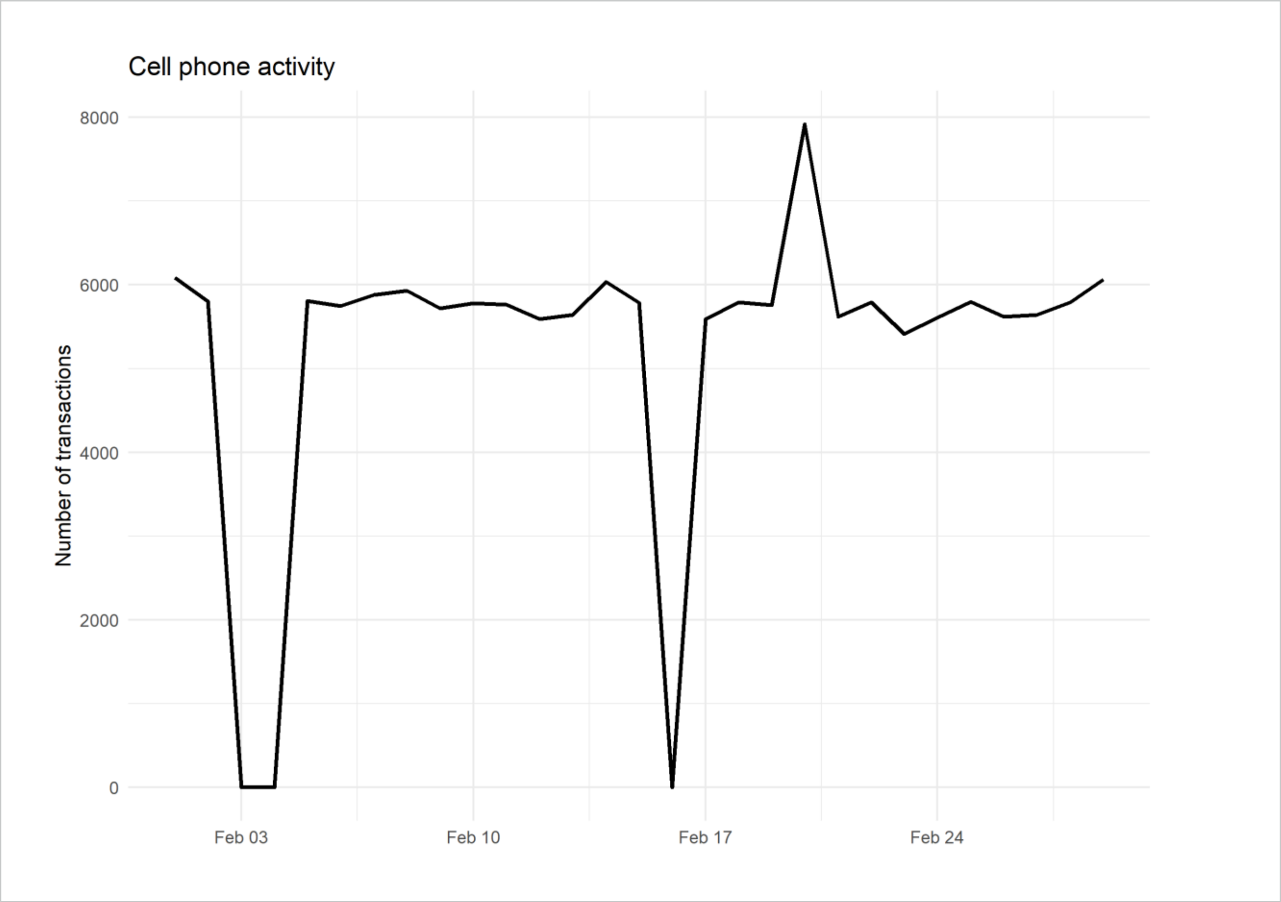
Check if data is internally consistent and with known facts or other data available such as

- Population size or MNO market share
- Number of subscribers is comparable to the number of calls.

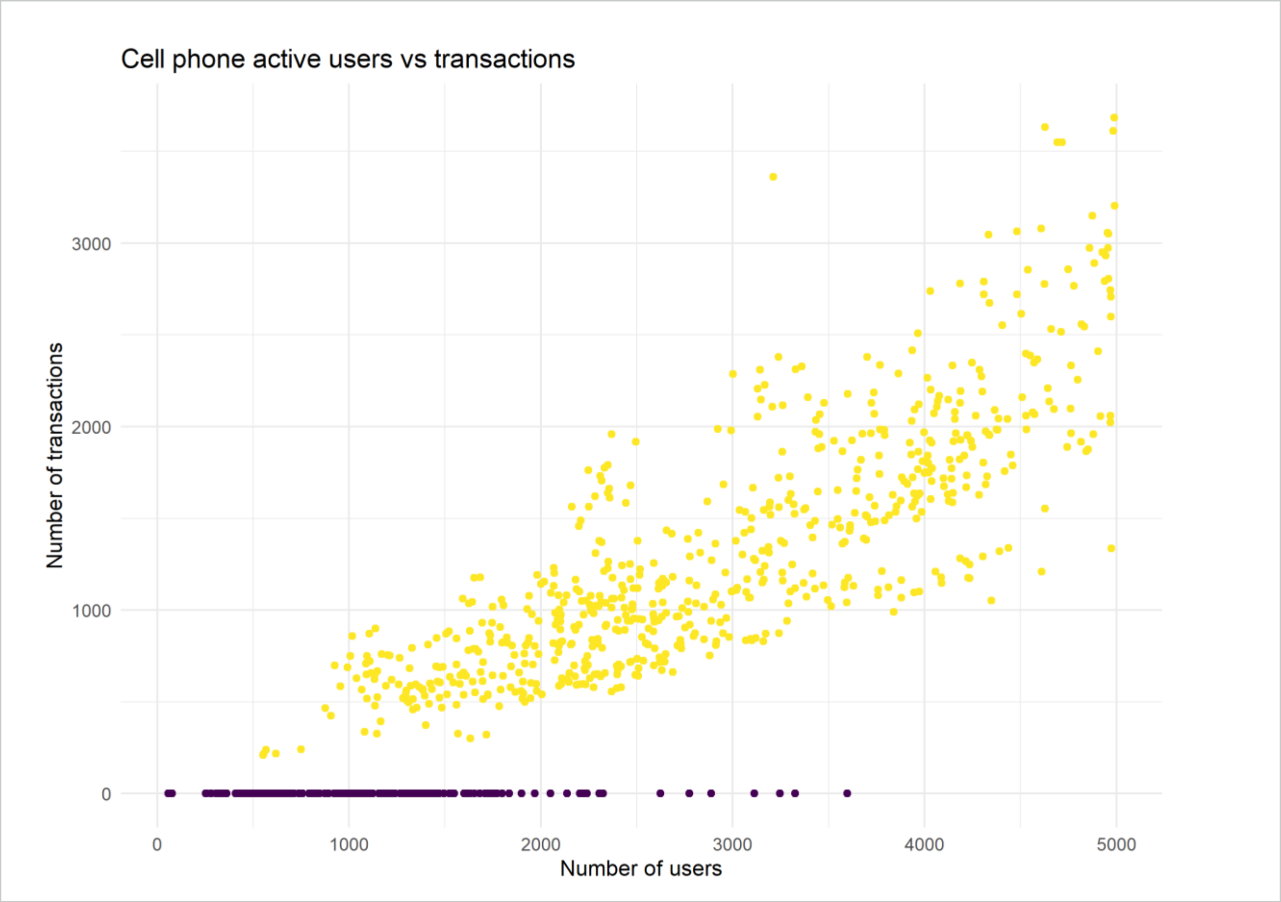
Anomalies

General checks for outliers or other general data issues:

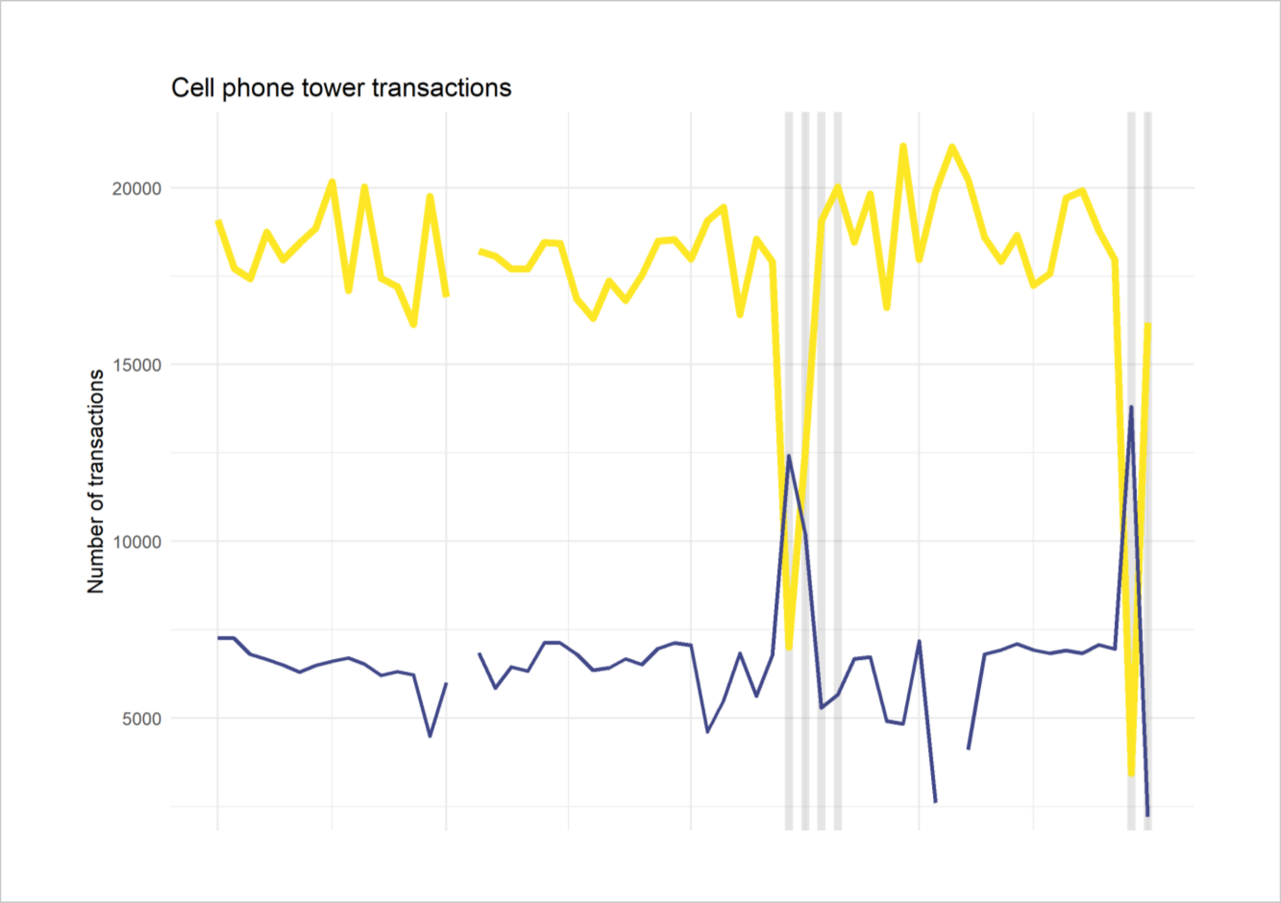
- Duplicated rows
- Sudden drops or spikes on specific regions.



Example: Completeness



Example: Consistency

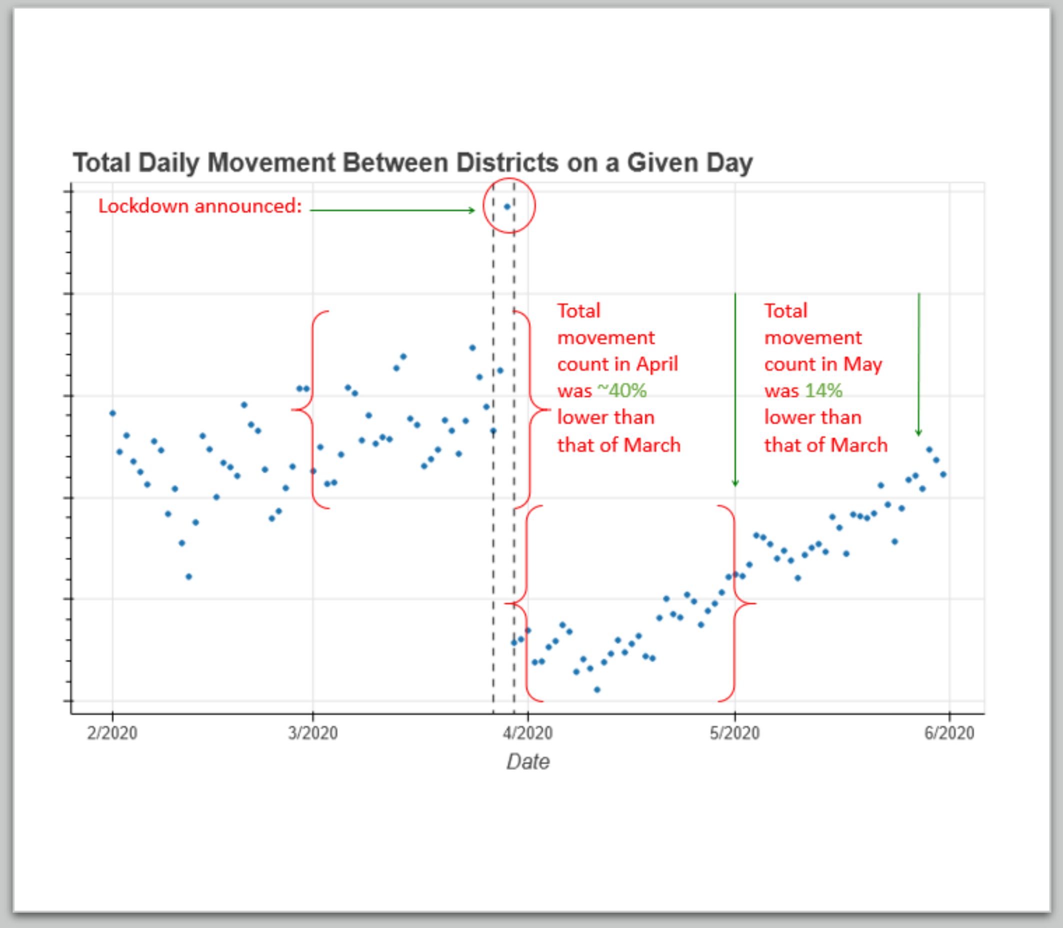


Example: Anomalies

Analysis and Interpretation

Analysis

- Once indicators are checked for quality, can be analyzed to inform policy
- Example: Change in mobility over time after a new policy



Interpretation

Data can be interpreted in different ways

For example, change in a variable can be measured differently, each with advantages and disadvantages

- Level comparison from previous day
- Percent change from previous day
- Percent change from a baseline (e.g month before COVID19 lockdown)
  - Definition of baseline can also vary:
    - Average across previous month
    - Average by day of week
    - Average by weekend vs weekday
- Z-score (avg adjusted for standard deviation)

Example:

	day 1	day 2	day 3	day 4	day 5
Travelers to district A	35	15	40	20	40
Travelers to district B	2500	3000	2700	2500	4000

Indicator	Change district A	Change district B
Level change from previous day	20	1500
% change from previous day	100%	75%
% change from baseline average	45%	49%
Z-score	1.2	6.5

Example:

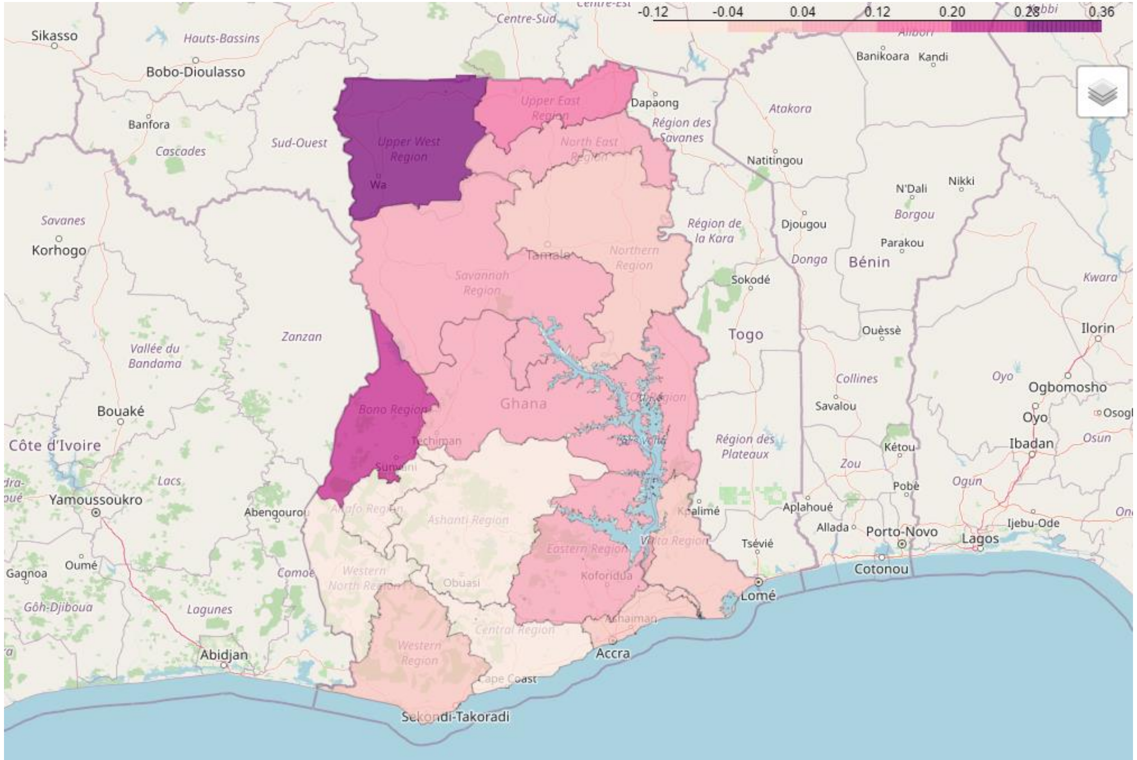
Baseline

	day 1	day 2	day 3	day 4	day 5
Travelers to district A	35	15	40	20	40
Travelers to district B	2500	3000	2700	2500	4000

Indicator	Change district A	Change district B
Level change from previous day	20	1500
% change from previous day	100%	75%
% change from baseline average	45%	49%
Z-score	1.2	6.5

Visualization

Visualization: Choropleth



Practical exercises

Instructions

1. The following files have been shared over e-mail:
   - admin1.geojson
   - movements_per_day.csv
   - subscribers_per_day.csv
   - transactions_per_day.csv
2. Use the link below to access the exercises notebook: https://colab.research.google.com/github/LeonardoViotti/cdr-training/blob/main/notebooks/aggregated-cdr-analysis.ipynb