

What is mobile phone data?

Overview of data generated by mobile communication technologies

Siim Esko

Positium

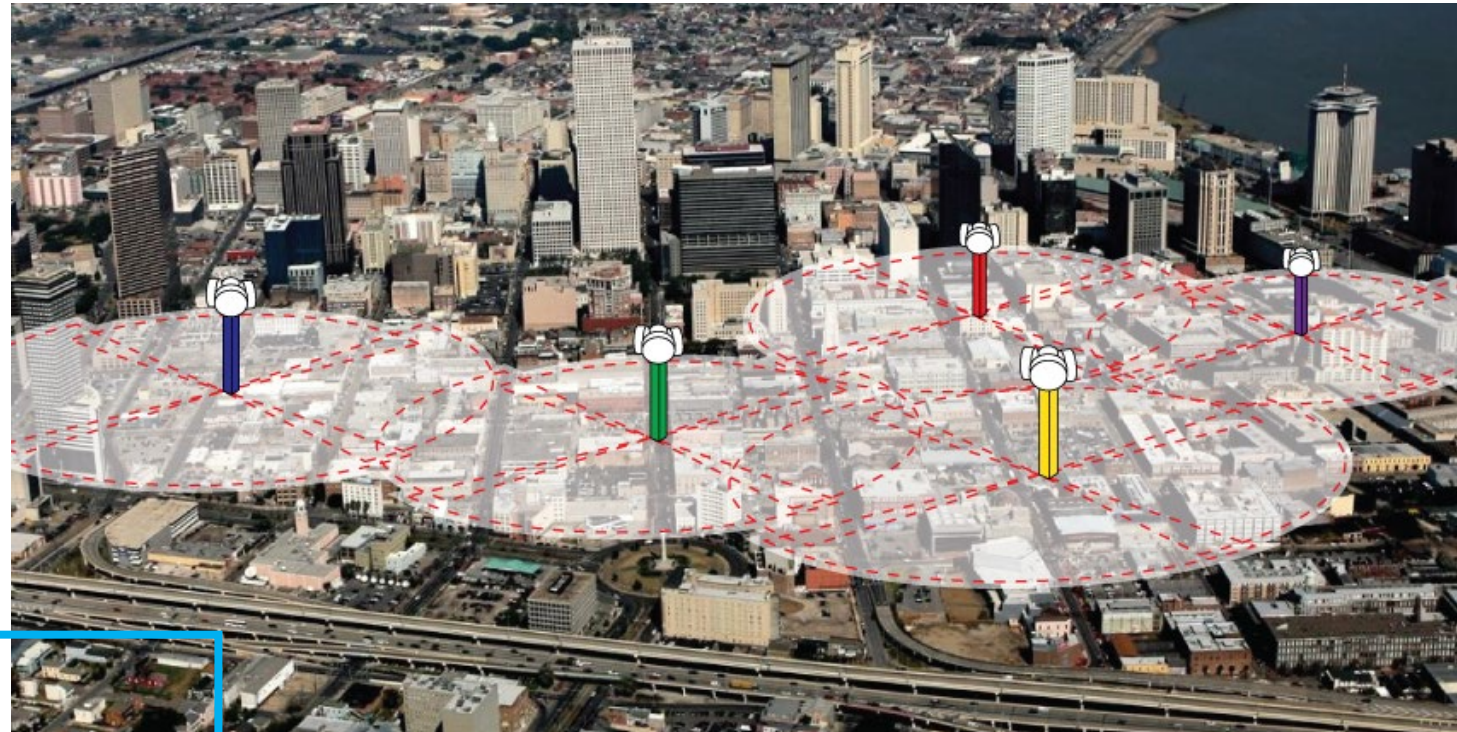
Estonia

@positium

www.positium.com

Mobile Positioning Data

- Active positioning data
 - Mobile Positioning System
 - App based positioning
- Passive positioning data
 - CDR's, IPDR's
 - Location area update, signalling data

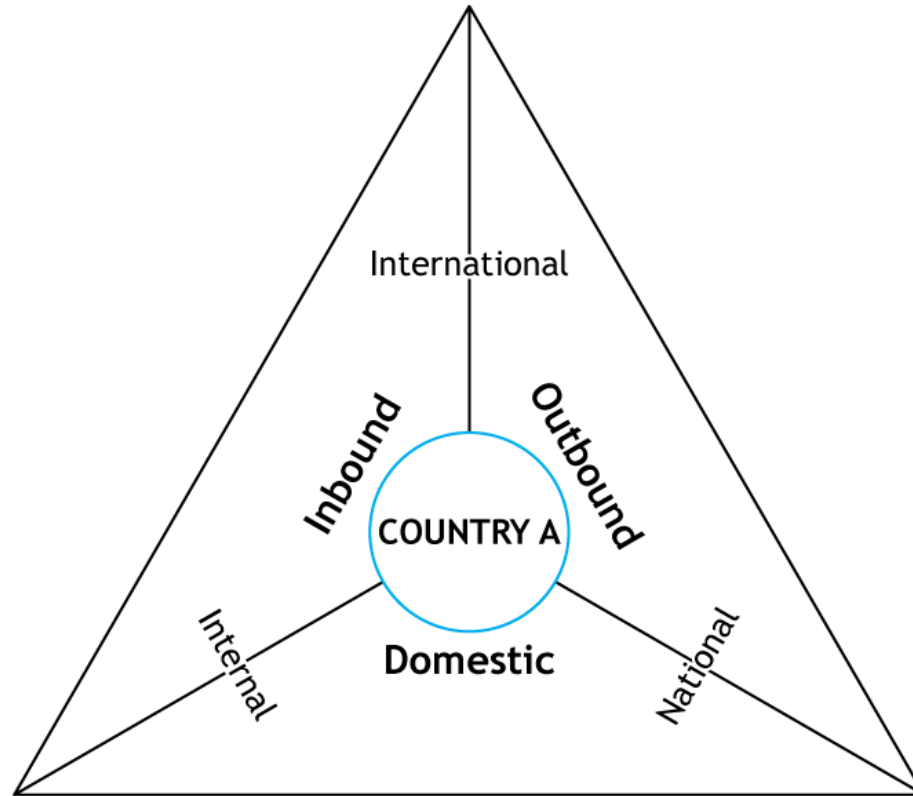


Mobile Positioning Data

ID	TIME	CELL_ID
2457971165	2017-07-17 04:06:59	63256
5011924318	2017-07-17 04:07:02	89772
2559837568	2017-07-17 04:07:03	45783
2859353729	2017-07-17 04:07:05	77674
5147064428	2017-07-17 04:07:07	24178
1747555920	2017-07-17 04:07:10	22307
1876419636	2017-07-17 04:07:11	80423
3083426241	2017-07-17 04:07:14	29019
6314522758	2017-07-17 04:07:17	32902
2502448573	2017-07-17 04:07:18	89499
4814713995	2017-07-17 04:07:21	60114
5557657815	2017-07-17 04:07:22	25312
6690183382	2017-07-17 04:07:23	12267
1195773102	2017-07-17 04:07:24	20509
3406038930	2017-07-17 04:07:26	28942
5513563945	2017-07-17 04:07:27	75770
6686956736	2017-07-17 04:07:30	28760
8313088852	2017-07-17 04:07:31	20824
6820801488	2017-07-17 04:07:32	66166
7354861093	2017-07-17 04:07:35	61721
6436449746	2017-07-17 04:07:37	98660
2324932872	2017-07-17 04:07:40	77357
5960310590	2017-07-17 04:07:42	19888
5036631668	2017-07-17 04:07:44	40159
7115373195	2017-07-17 04:07:45	83692
7374888775	2017-07-17 04:07:47	91126
4883201171	2017-07-17 04:07:49	50219



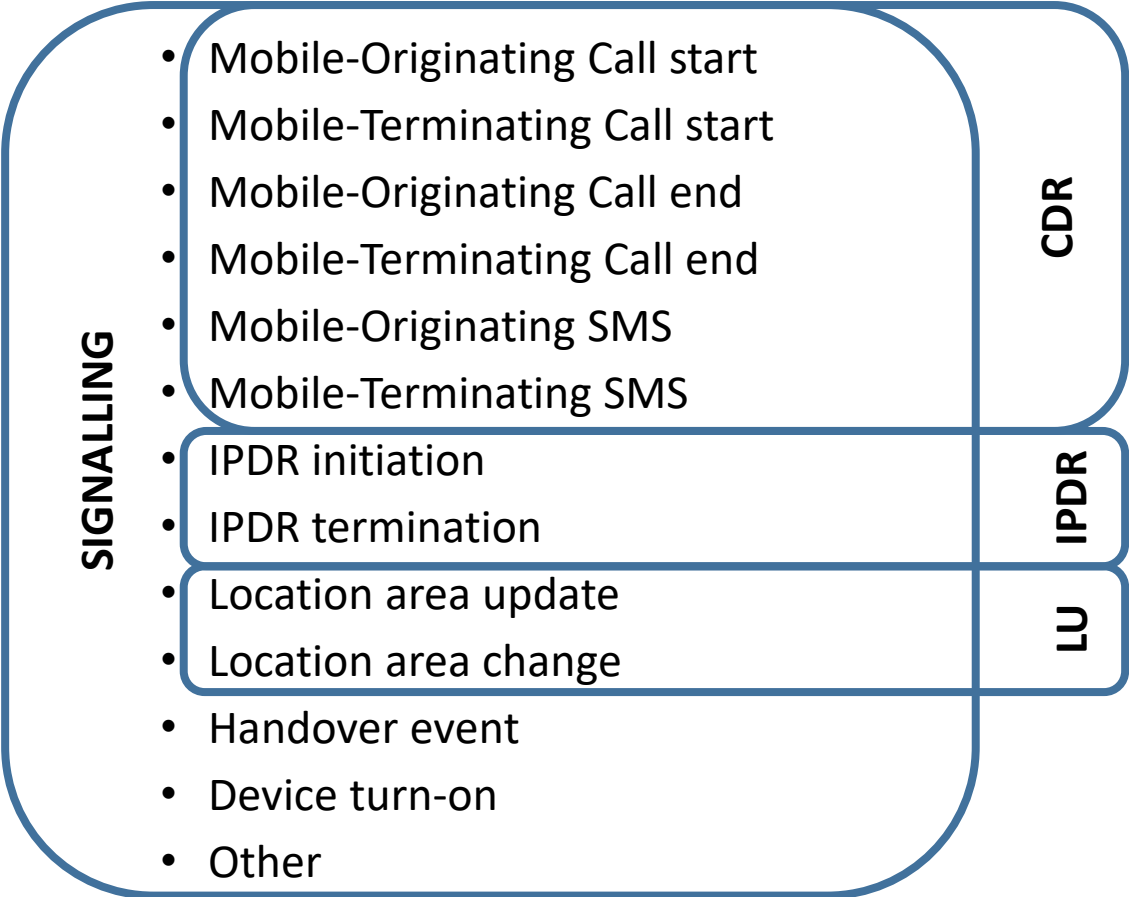
Forms of Data



Types of Data

- Call Detail Records (CDR)
 - Outgoing
 - Calls, SMS
 - Incoming
 - Calls, SMS
- Internet Protocol Detail Records (IPDR)
 - Mobile internet usage
- Location area update
 - LA change due to movement
 - No activity > 2hr => LA update
- Signalling, Abis, Probe data
 - Virtually any activity that involves communication between the device and the network
 - Signalling includes all types above + additional – handovers, phone switch on, etc.

Data Events



Data types include different events

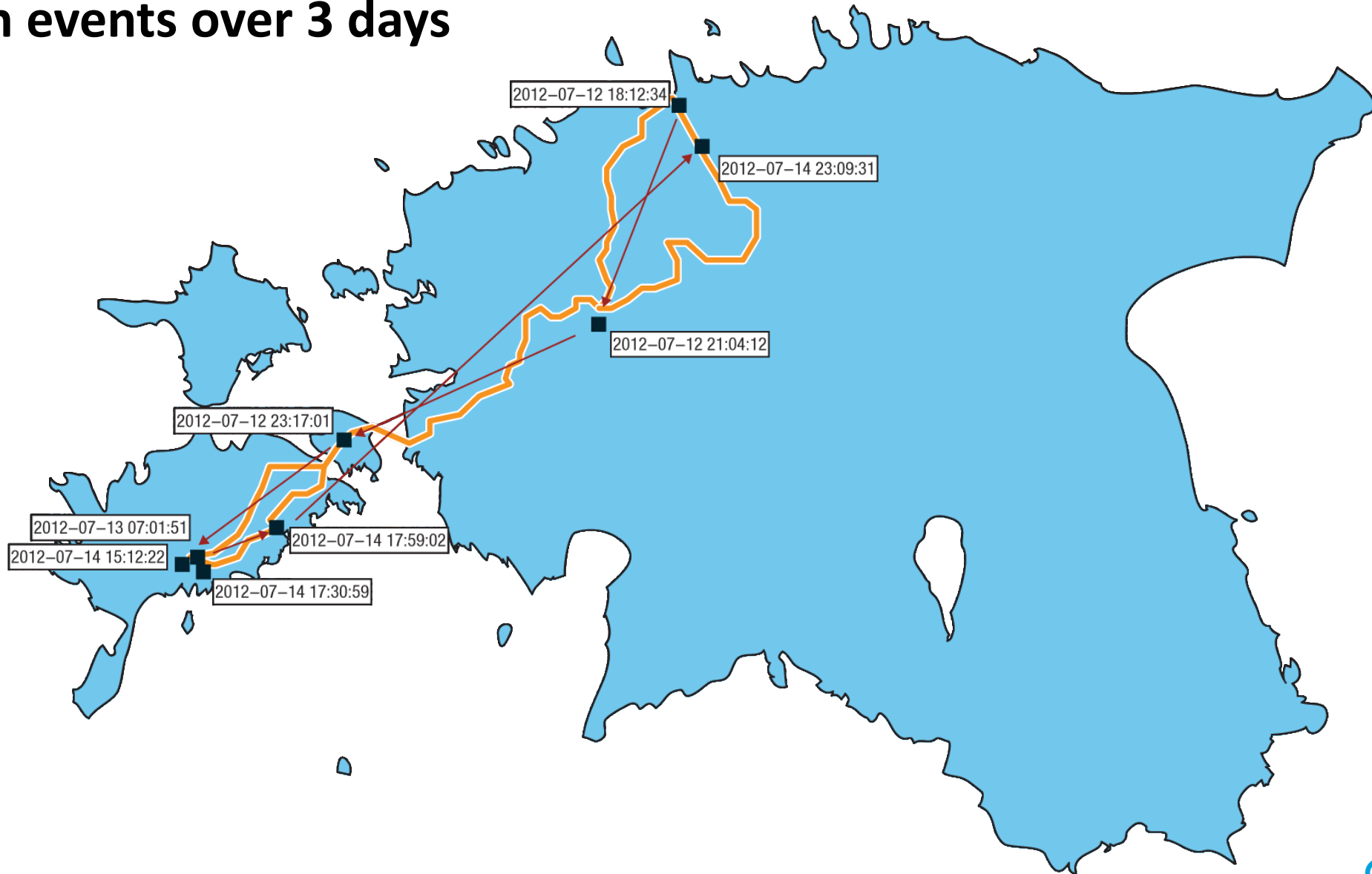
Main 3 Attributes in Raw Data



ID	Timestamp	Cell_id	MCC
IMSI based (pseudonymised) ID of the subscriber	Time of the event	Location of the event <ul style="list-style-type: none">• Cell_id reference• Geographical coordinates• Geographical coverage area	Mobile Country Code Necessary to distinguish domestic from inbound roaming

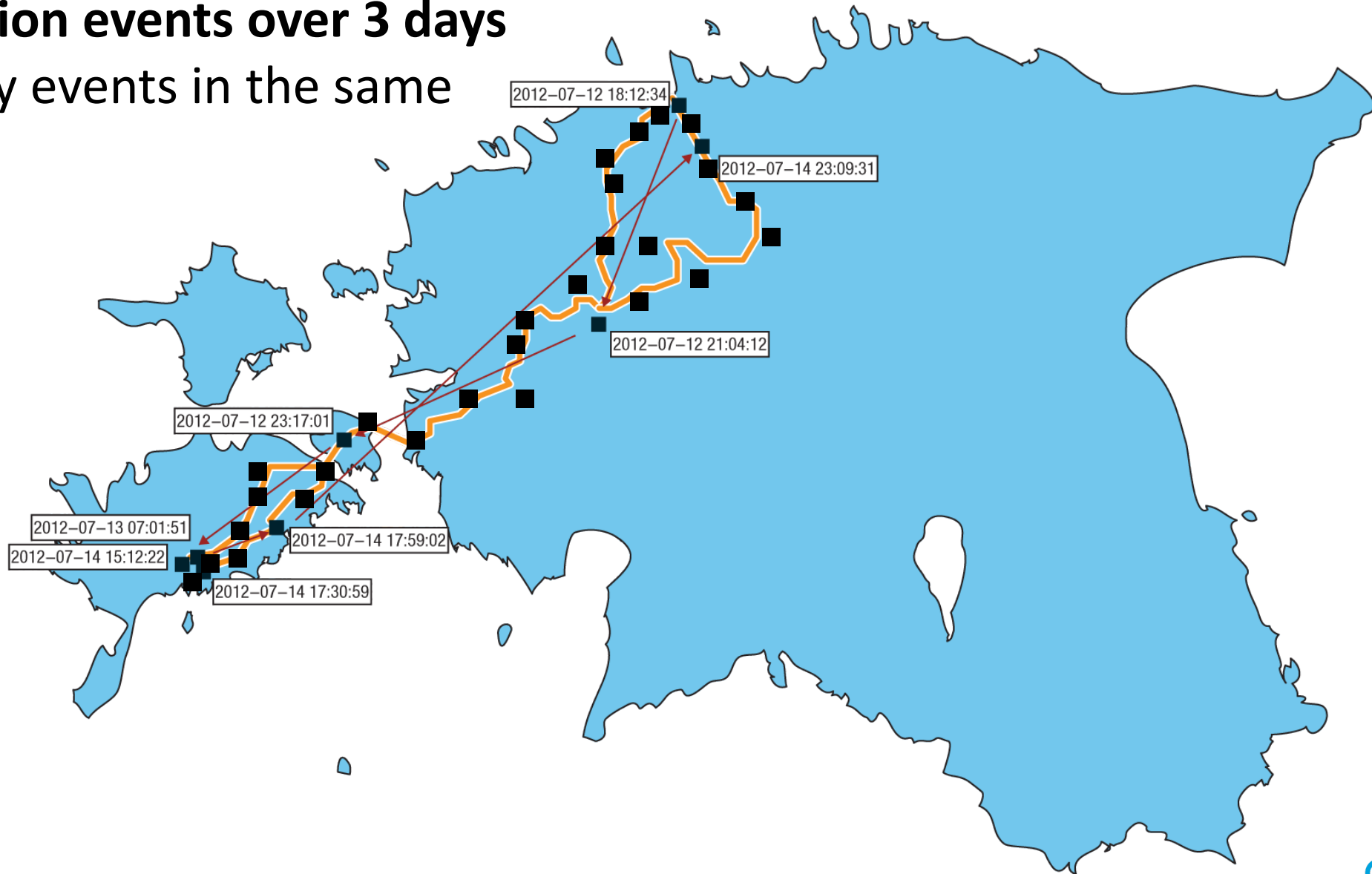
Example of CDR data

8 location events over 3 days



Example of Signalling Data

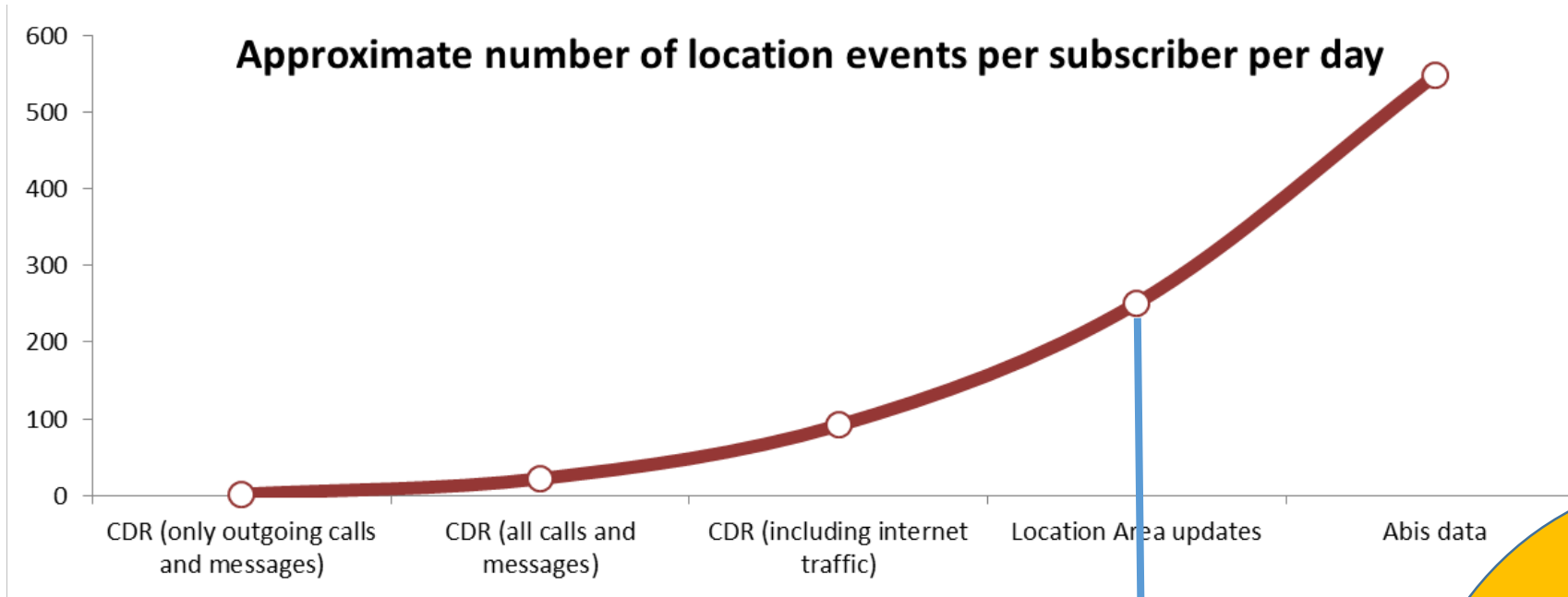
289 location events over 3 days
(but many events in the same location)



Availability of Different Types of Data

Data type	Availability in MNO	Approx. Avg number of records per subscriber per day
CDR outgoing	Always stored (billing)	2-3
CDR incoming	If MNO has decided to store	2-3
IPDR	If MNO has decided to store	0...200
LA update	If MNO has decided to store	12
Signalling	Requires additional network hardware installation + vast storage amounts (Hadoop HDFS)	200...2000

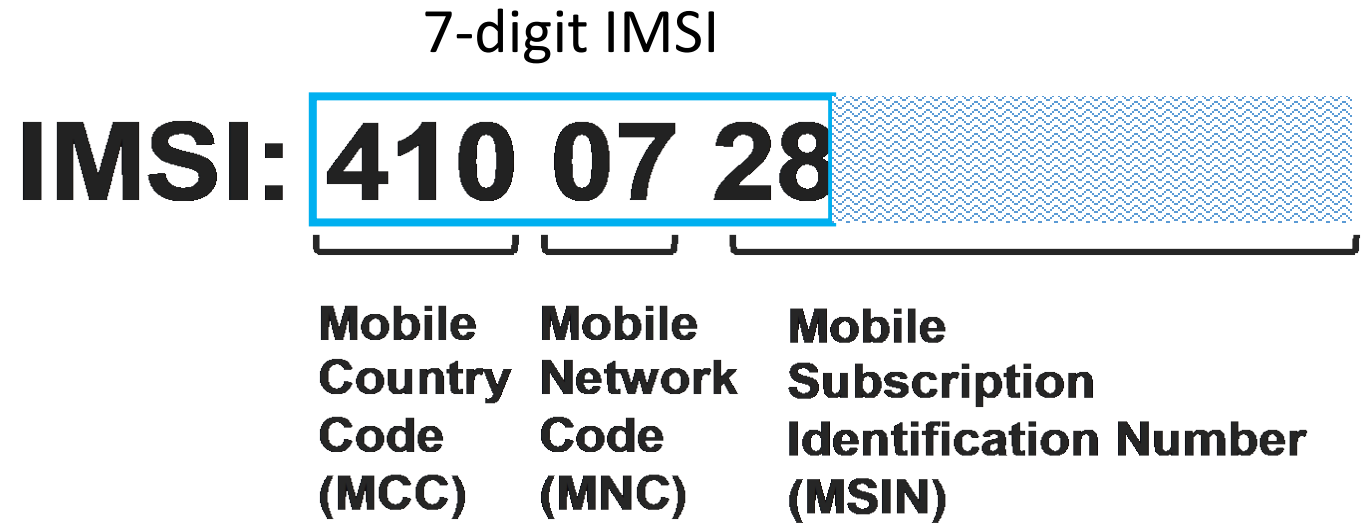
Data Size – 1 Year – Example Indonesia



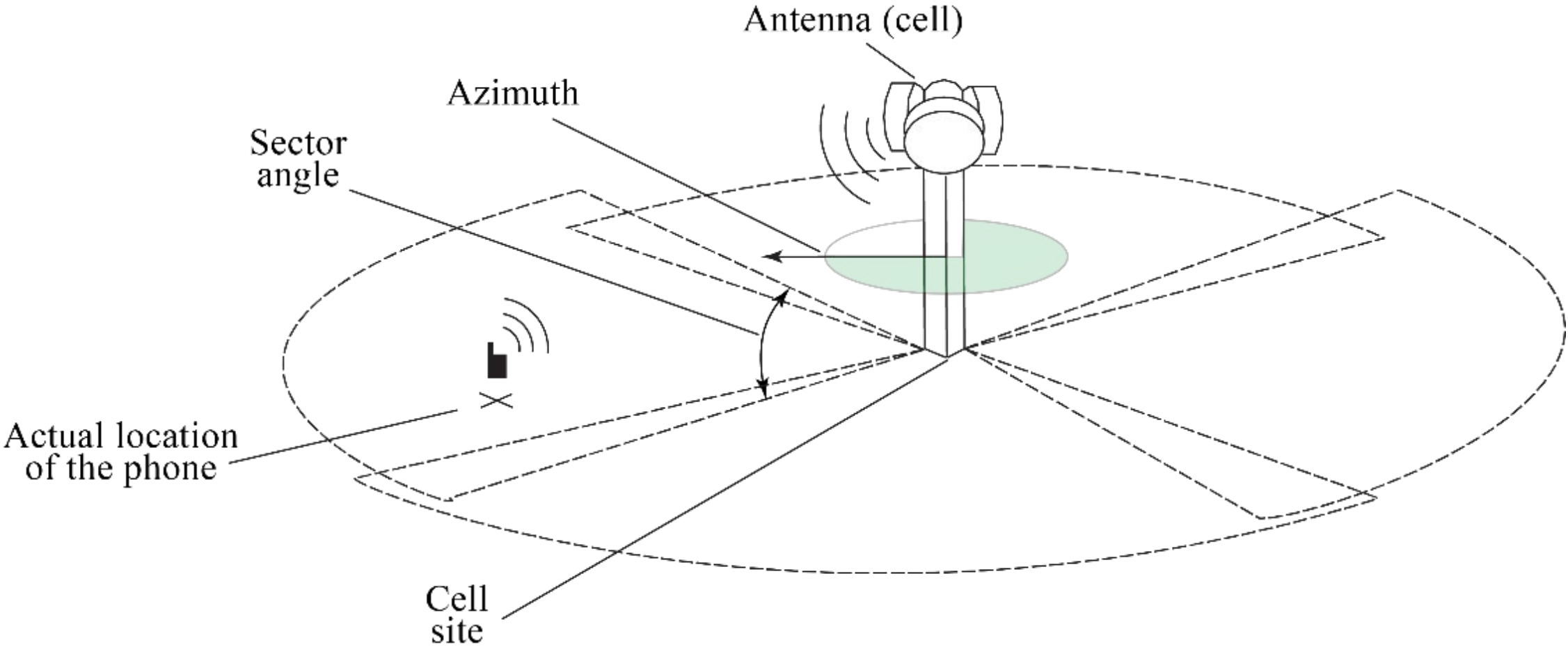
Data type	Subscribers	No of records	Volume
Inbound	12M	1B	0.3 TB
Outbound	20M	1.5B	0.5 TB
Domestic	246M	1500B	500 TB

ID of the Subscriber

- Four possibilities of ID:
 - IMSI (=>MCC)
 - MSISDN (=>CC)
 - IMEI (=>TAC)
 - MNO customer ID
- IMSI is preferred ID as raw data
- Hashed into pseudonymous ID



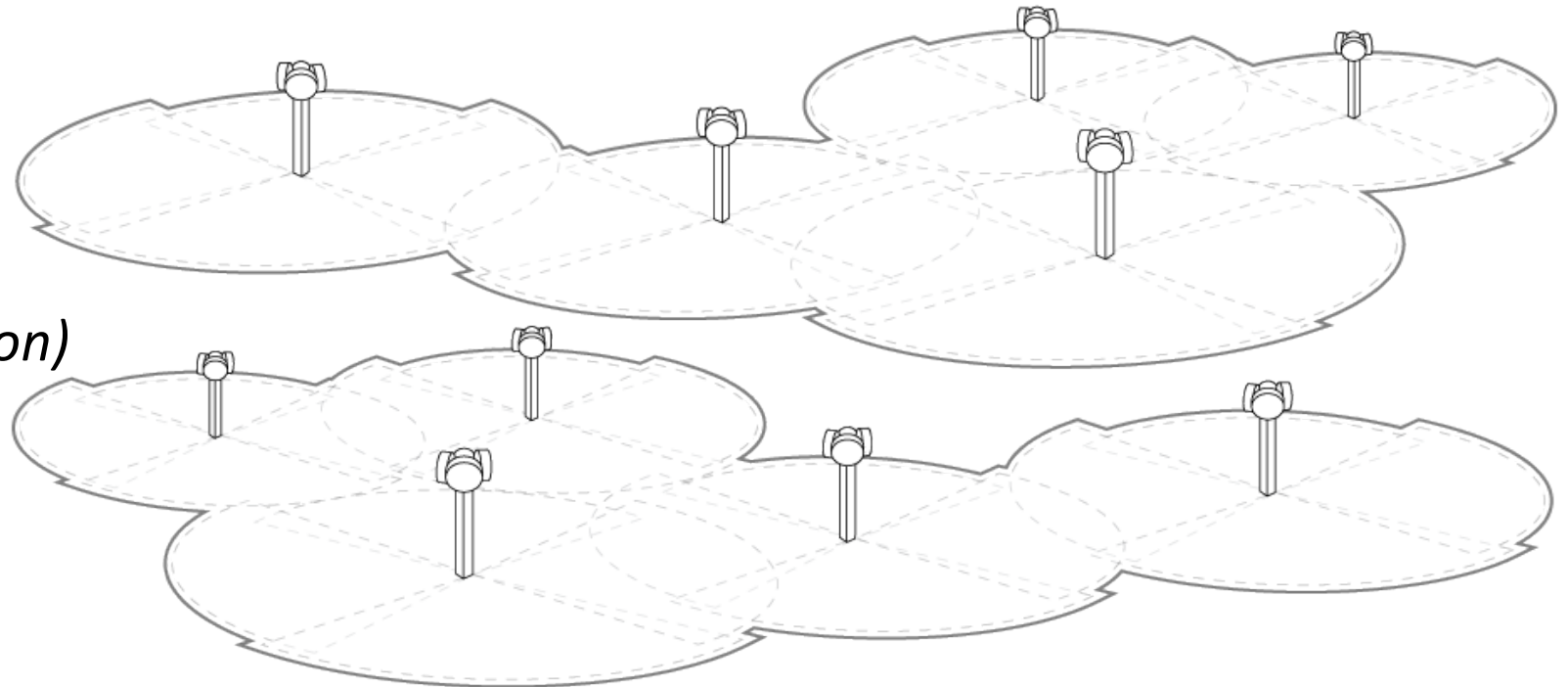
Mobile network antenna



Network Antennae Data

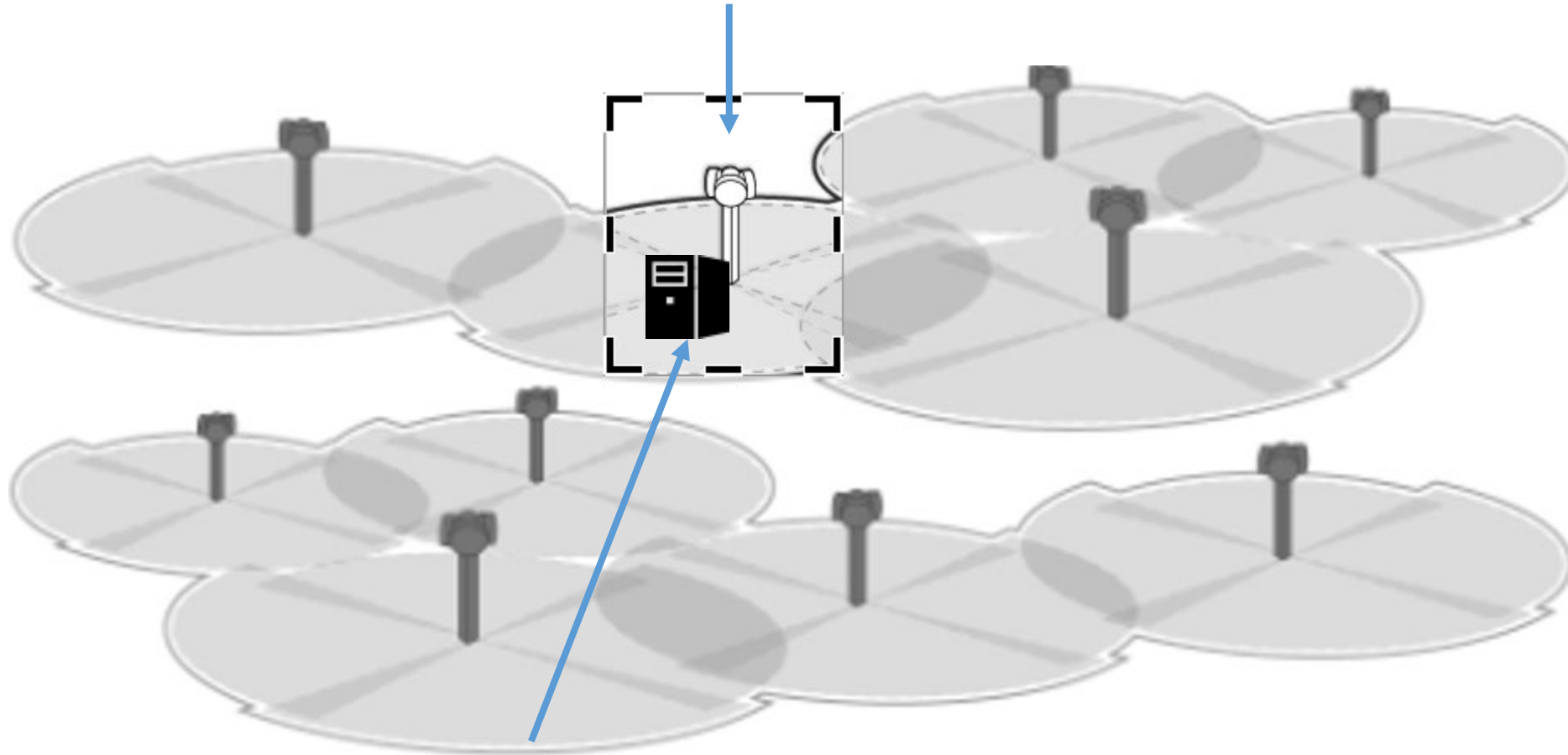
- Network antennae database includes:

- CELL_ID
- Type (2G, 3G, LTE)
- Location coordinates
 - Lat
 - Lon
- *Coverage area (polygon)*
- *Height of the cell*
- *Power (db)*
- *Azimuth*
- *Angle*



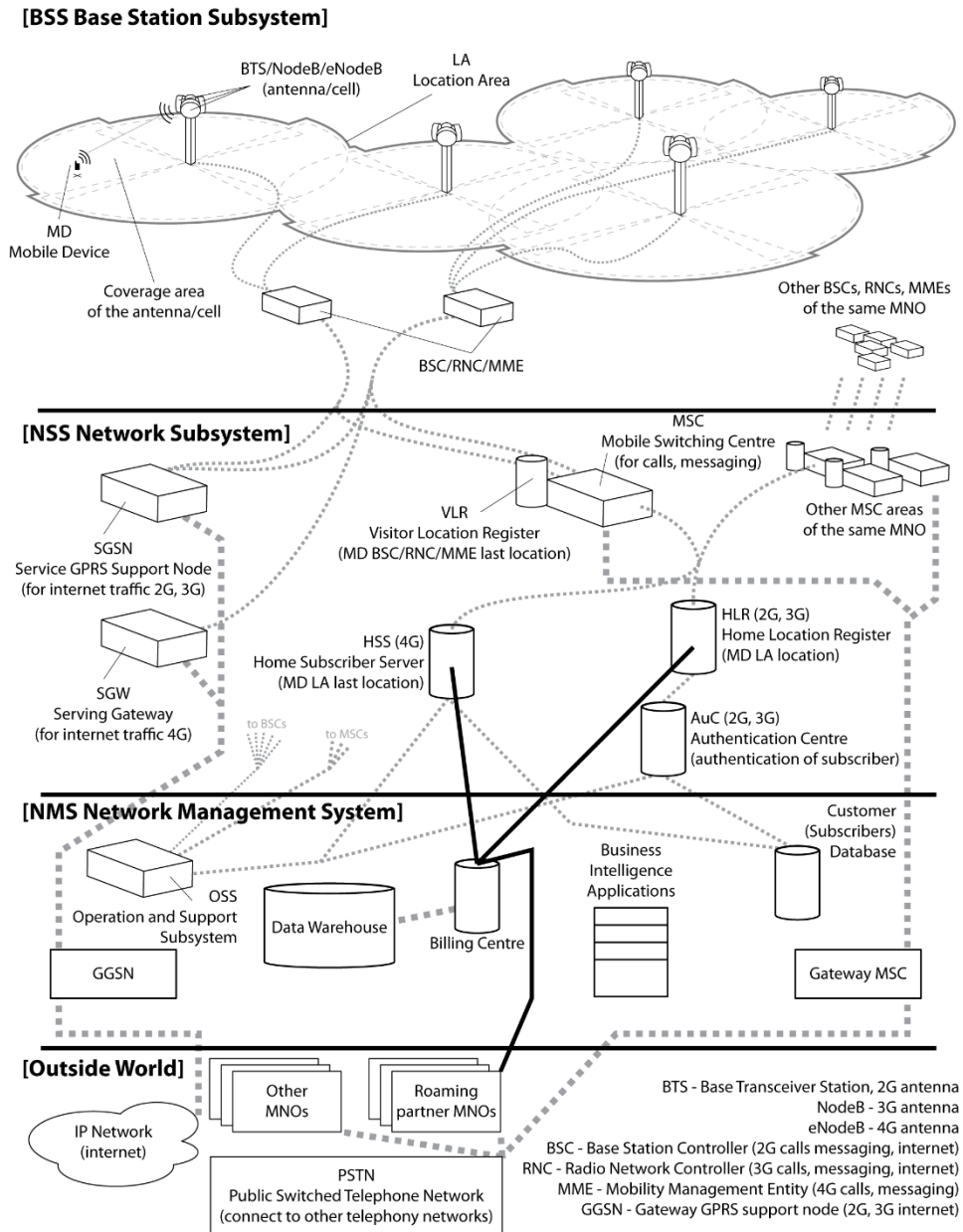
Base Station / Base Station Controller

There can be several cells in one location: E.g.: 6*2G, 2*3G, 1*LTE
One cell = one Base Station (BTS), also called node b (3G), eNB (LTE)

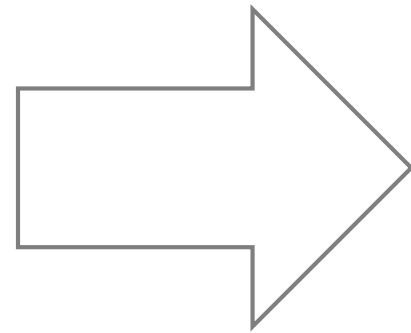


Base Station Controller (BSC) can control several Base Stations

Passive Mobile Positioning Data



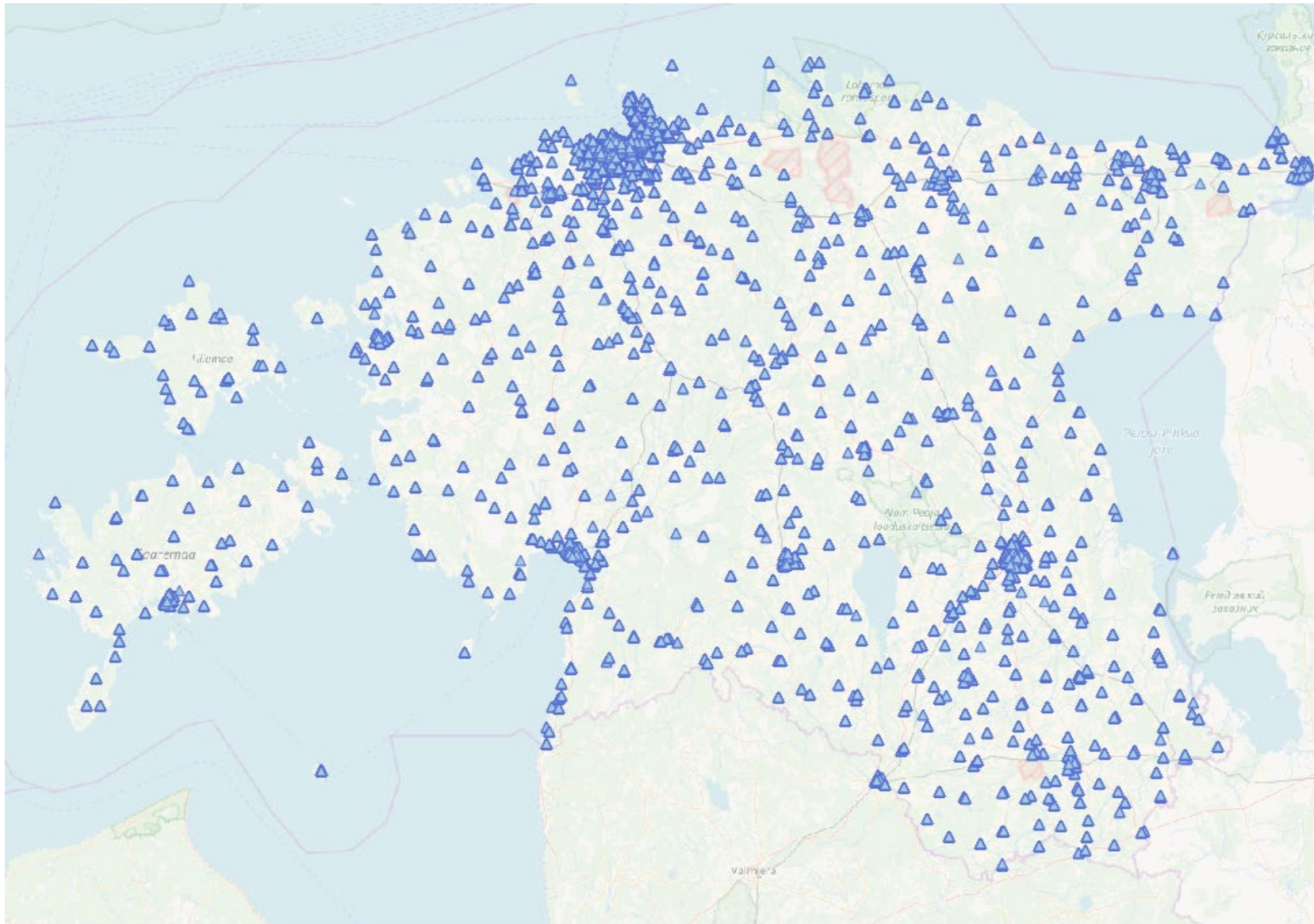
Historical residual data (log files, metadata) from mobile network operators' databases



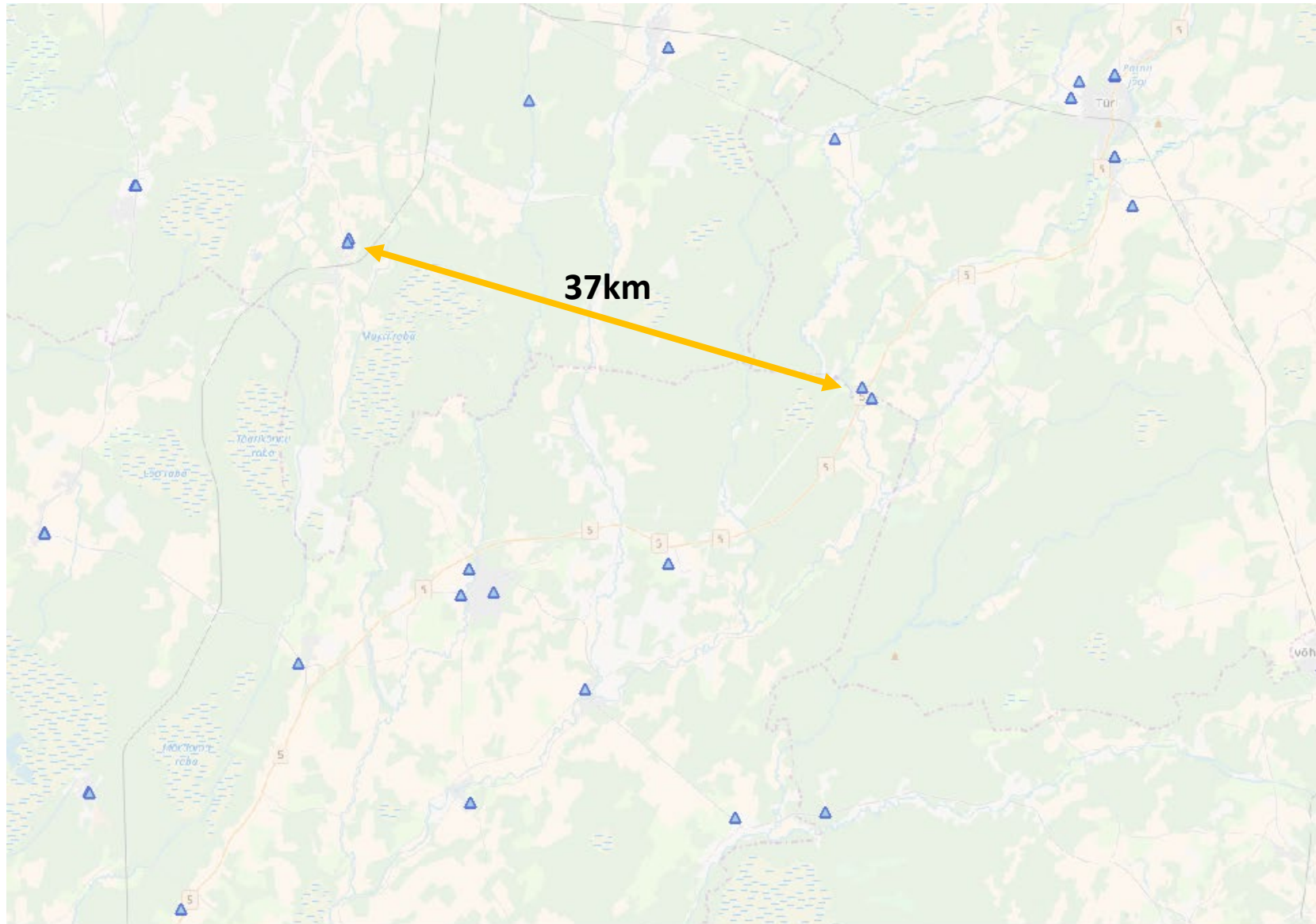
```
<?xml version="1.0"?>
<cdr>
  <id>b478fndh4dcdvjt</id>
  <start_time>2014-04-23 14:34:12</start_time>
  <start_epoch>1217529338</start_epoch>
  <start_uepoch>1217529338212616</start_uepoch>
  <answer_epoch>1217529338</answer_epoch>
  <answer_uepoch>1217529338452698</answer_uepoch>
  <end_epoch>1217529377</end_epoch>
  <end_uepoch>121752937795951</end_uepoch>
  <end_time>2014-04-23 14:35:14</end_time>
  <billsec>62</billsec>
  <success>1</success>
  <imsi>384845321548865</imsi>
  <type>call_out</type>
  <type_o>voice</type_o>
  <gen>2g</gen>
  <cgi>248487865145</cgi>
  <msisdn_rec>284872535485</msisdn_rec>
  <default_area_code>24807</default_area_code>
  <duration>00:01:02</duration>
  <record_type>125425</record_type>
  <accountcode>1160</accountcode>
  <user_context>default</user_context>
</cdr>
```


Antennae Distribution: Whole Country

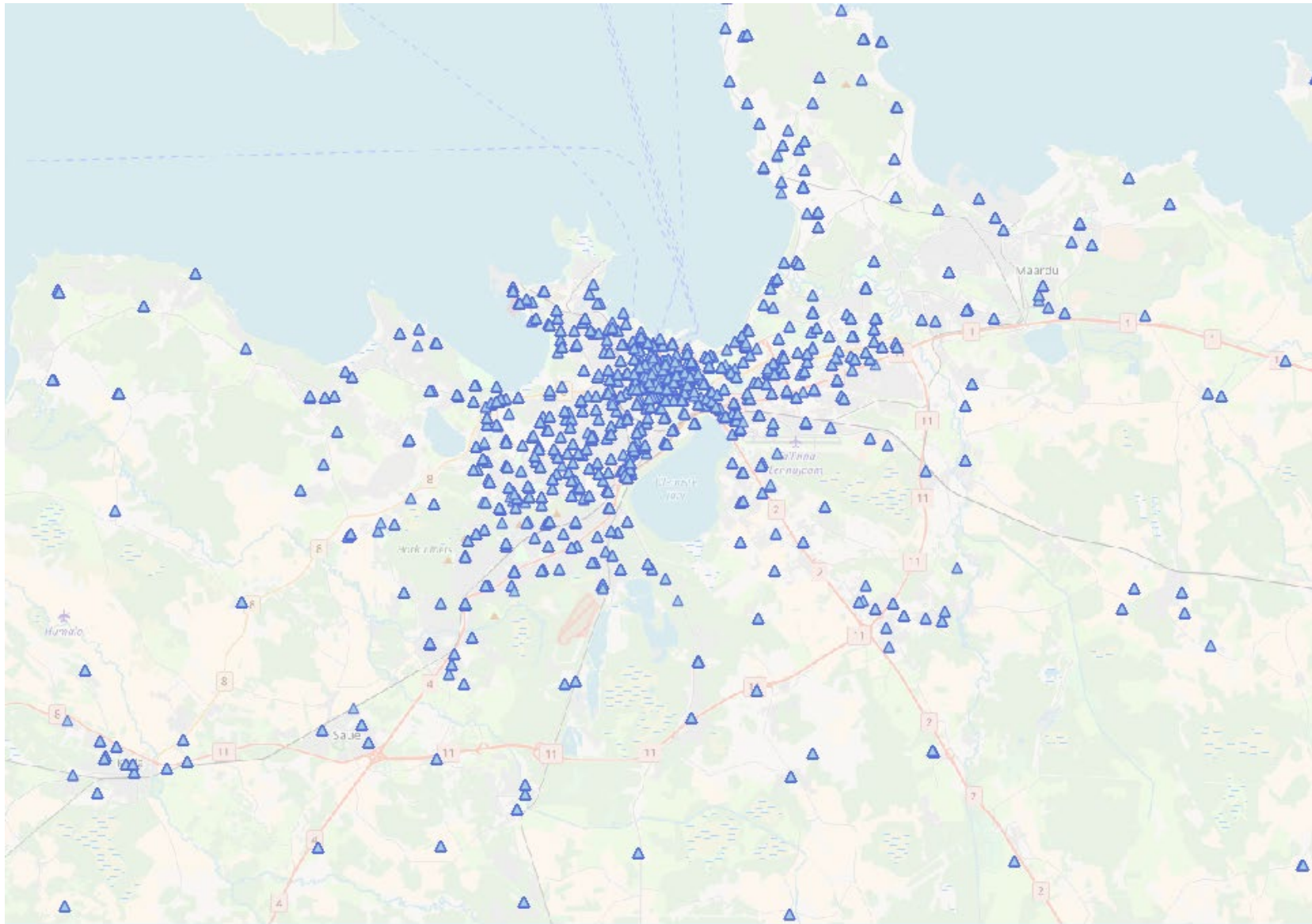
All cells/antennae in Estonia: 36,000



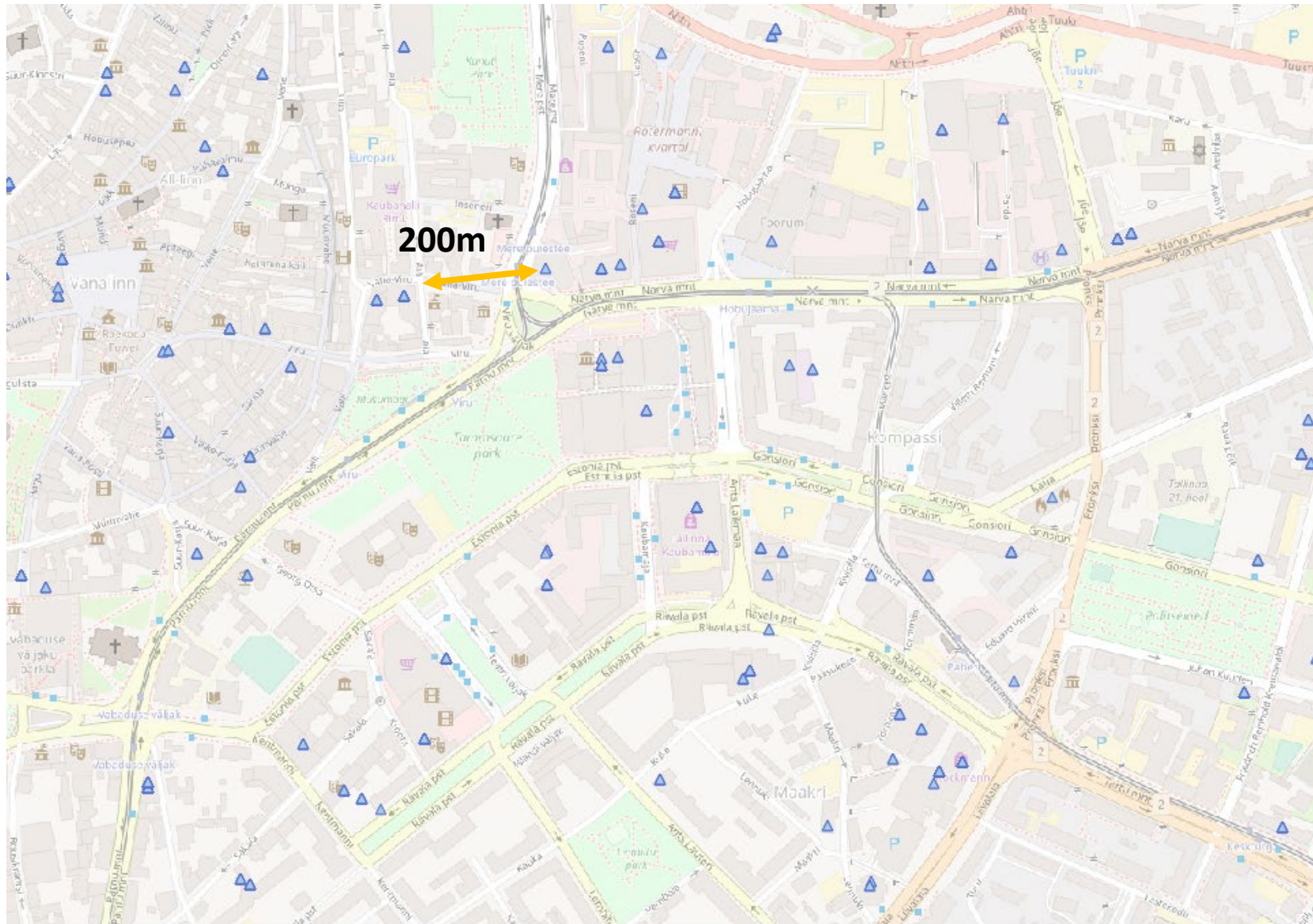
Antennae Distribution: Rural Areas



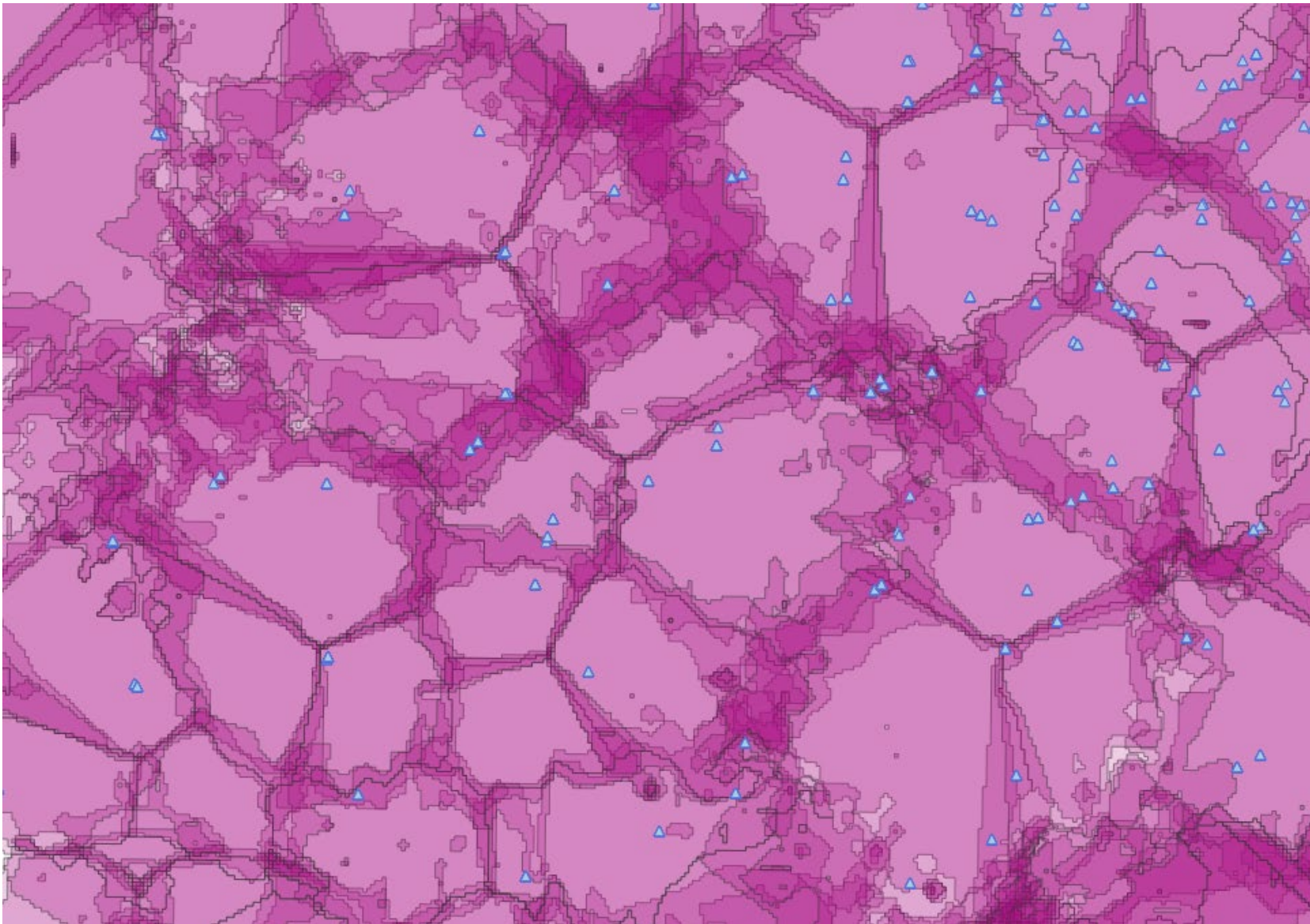
Antennae Distribution: Urban Areas



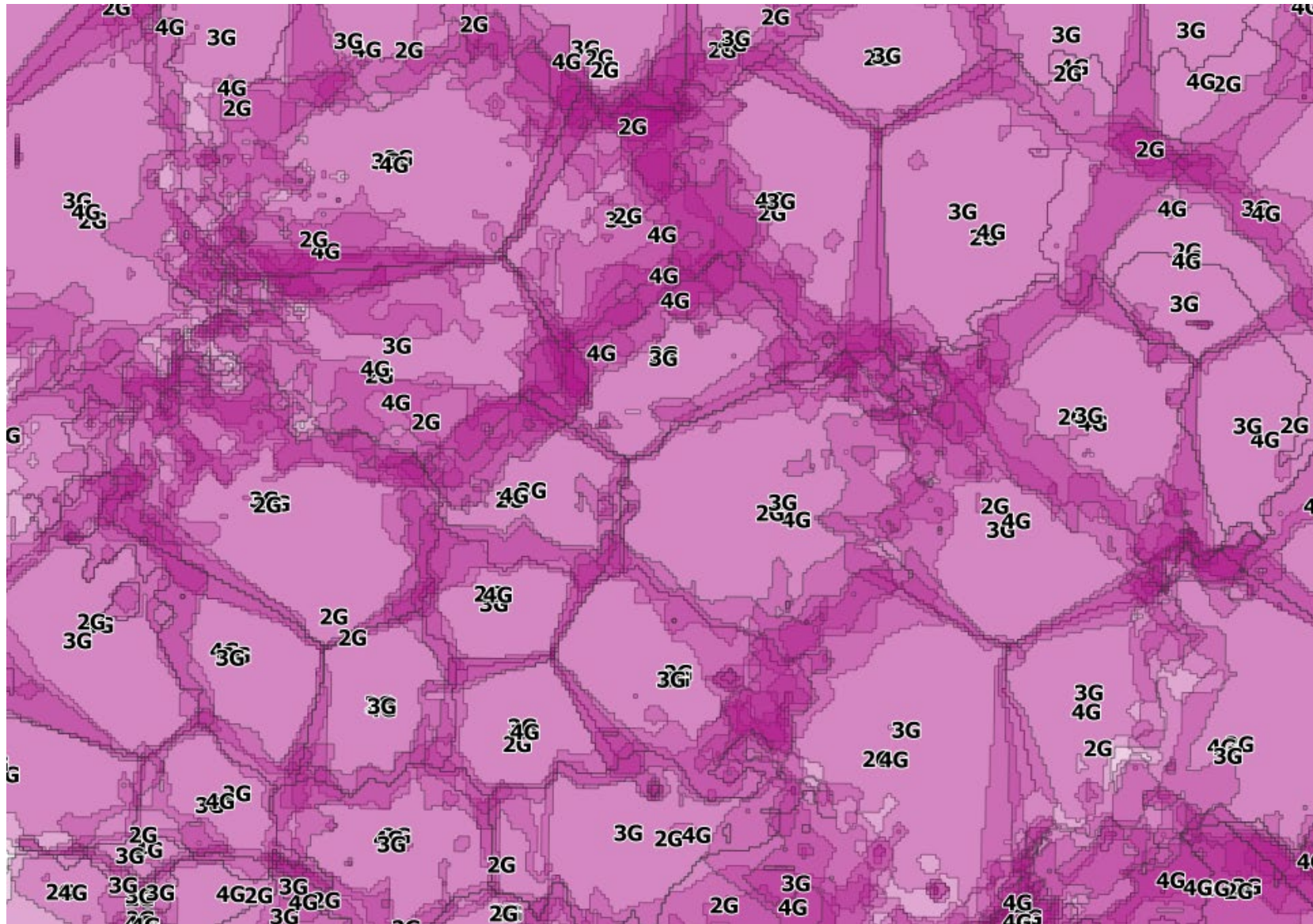
Antennae Distribution: Urban Areas



Antennae Coverage: Urban Areas



Antennae Coverage: Urban Areas



Network Antennae Data

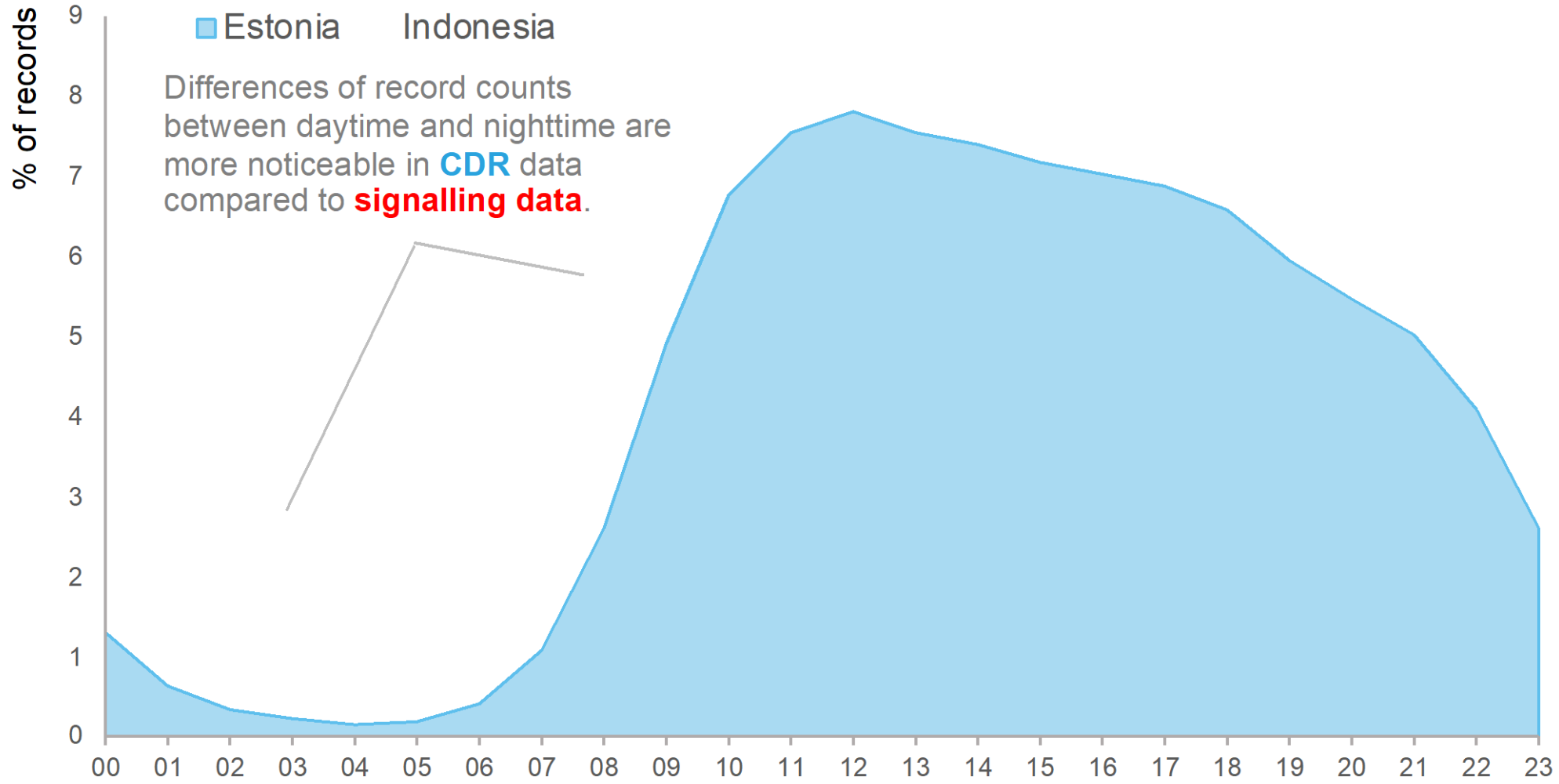
- Very often MNOs do not provide coverage areas, only coordinates (lat, lon)
- Even if they provide coverage areas, these are modelled and might be incorrect (but still better than just coordinates)
- Network antennae data should be checked for incorrect values (covered in topic 5)
- If you wish for results in high level spatial resolution, then modelling and interpolation is required (covered in topic 5)

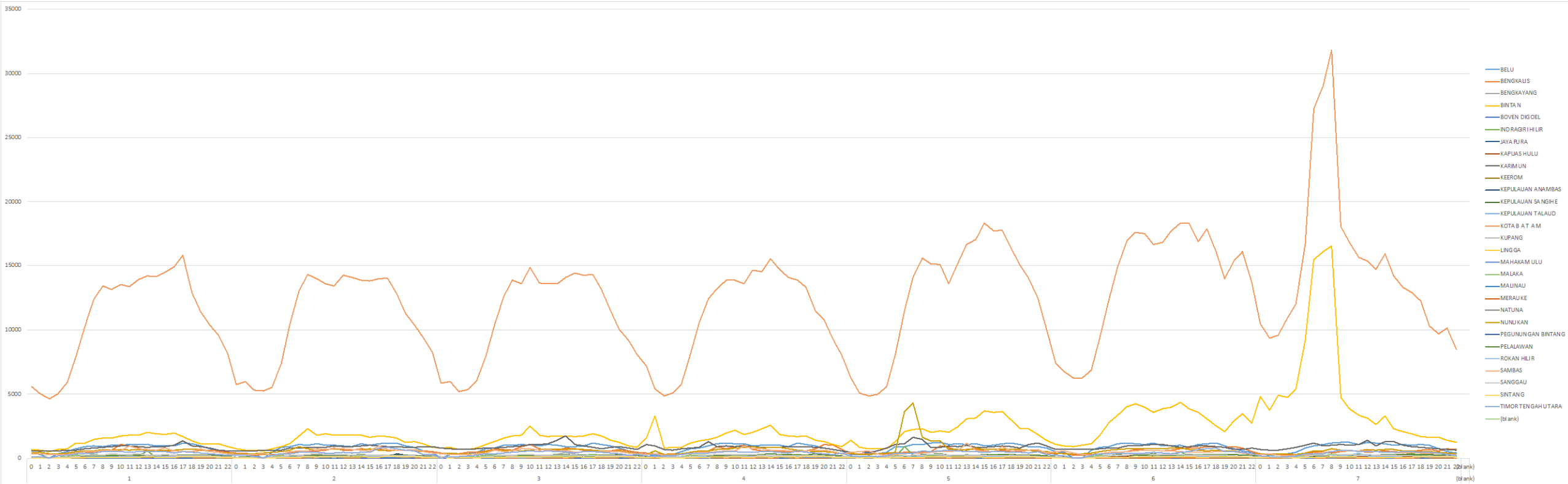
Estonia = CDR

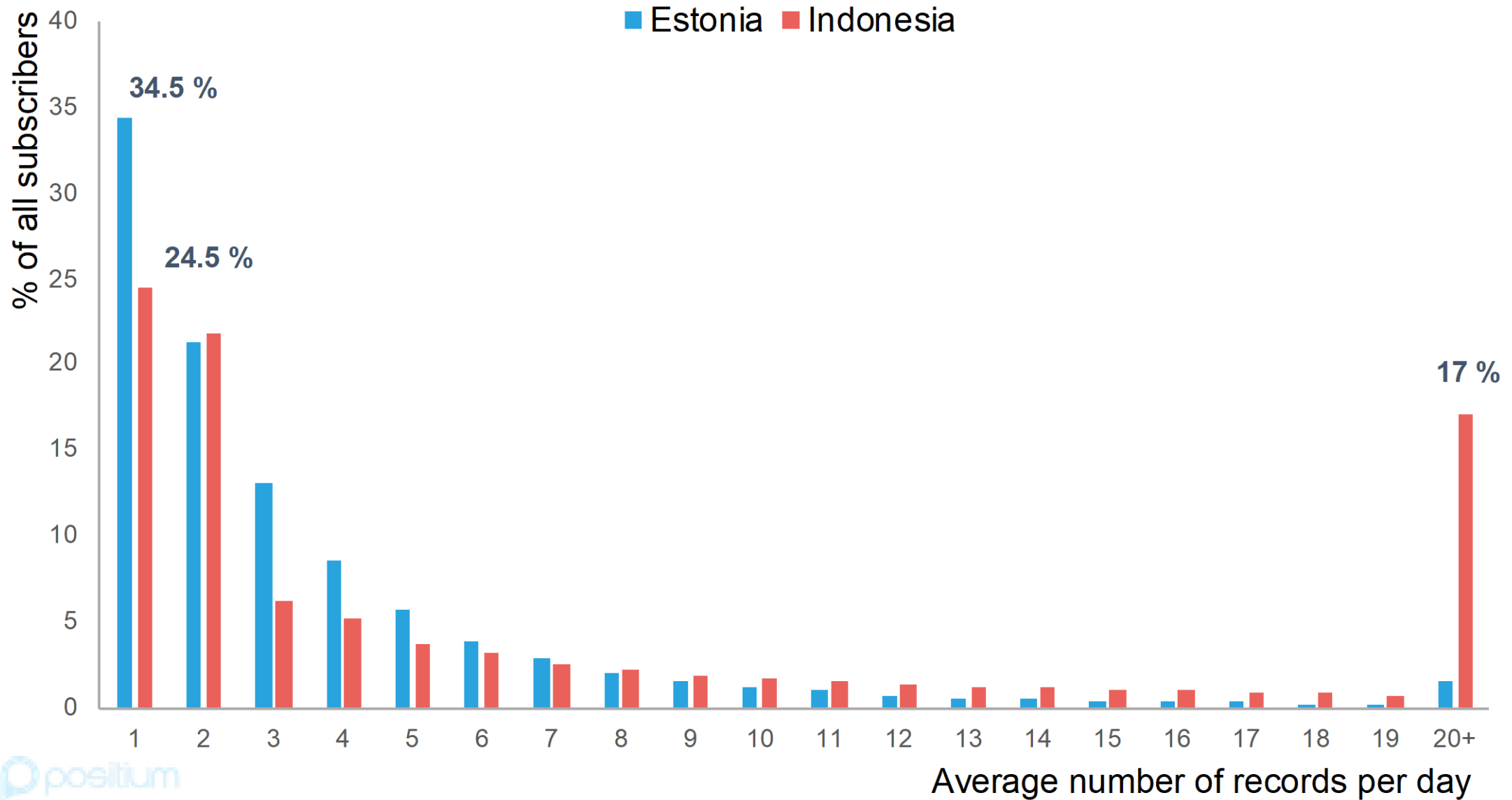
Indonesia = Signalling

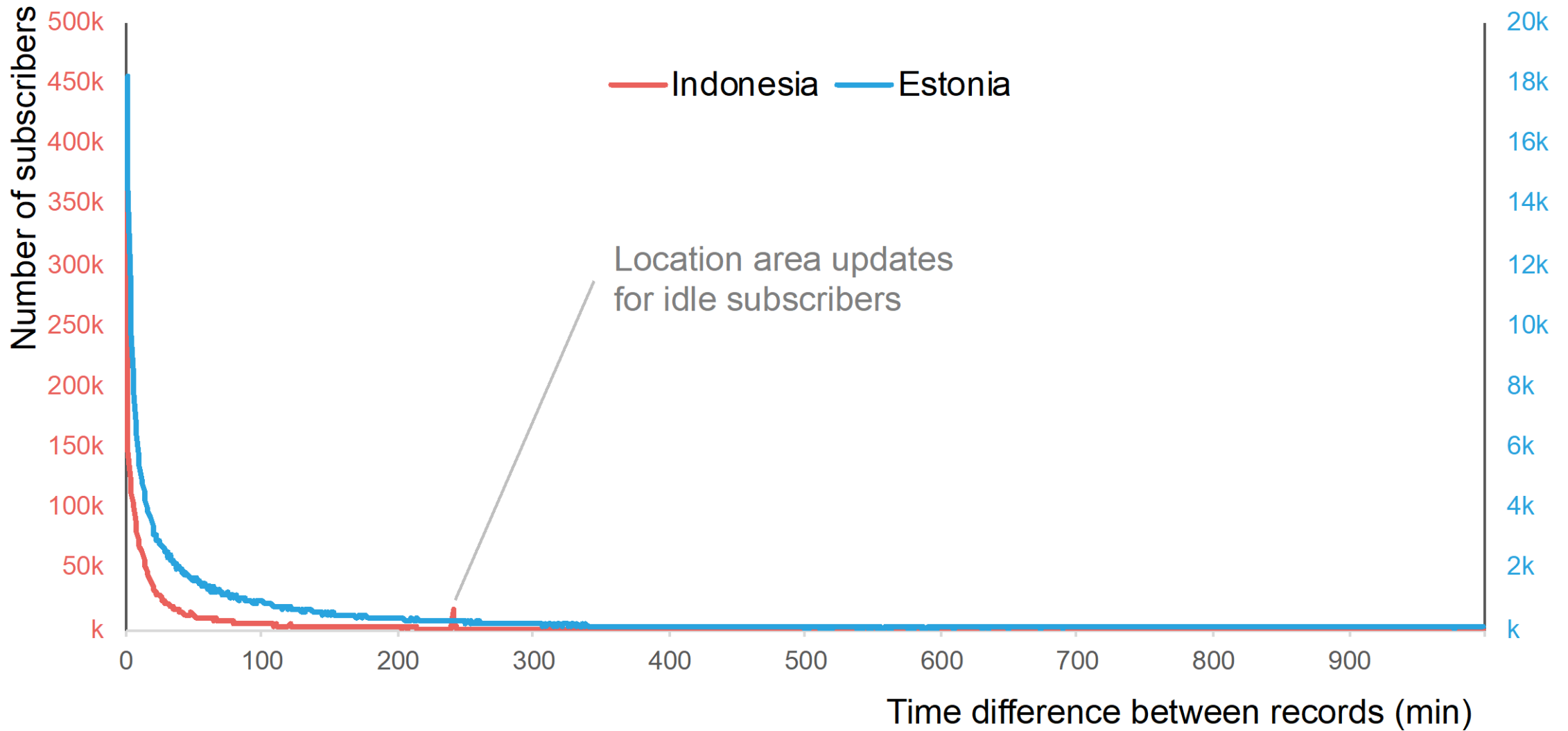
Georgia = CDR

Some Descriptive Statistics – Estonia vs Indonesia



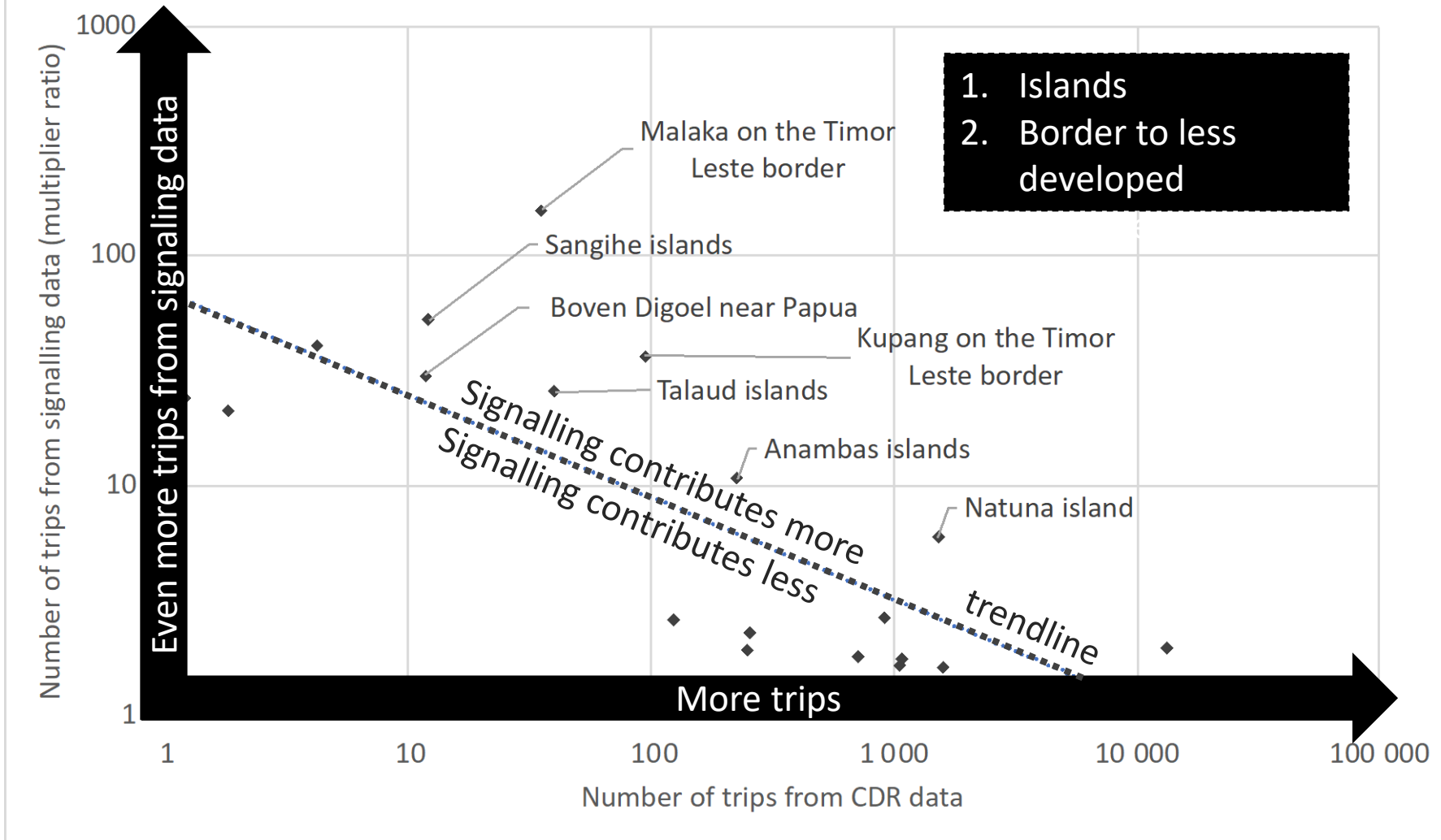




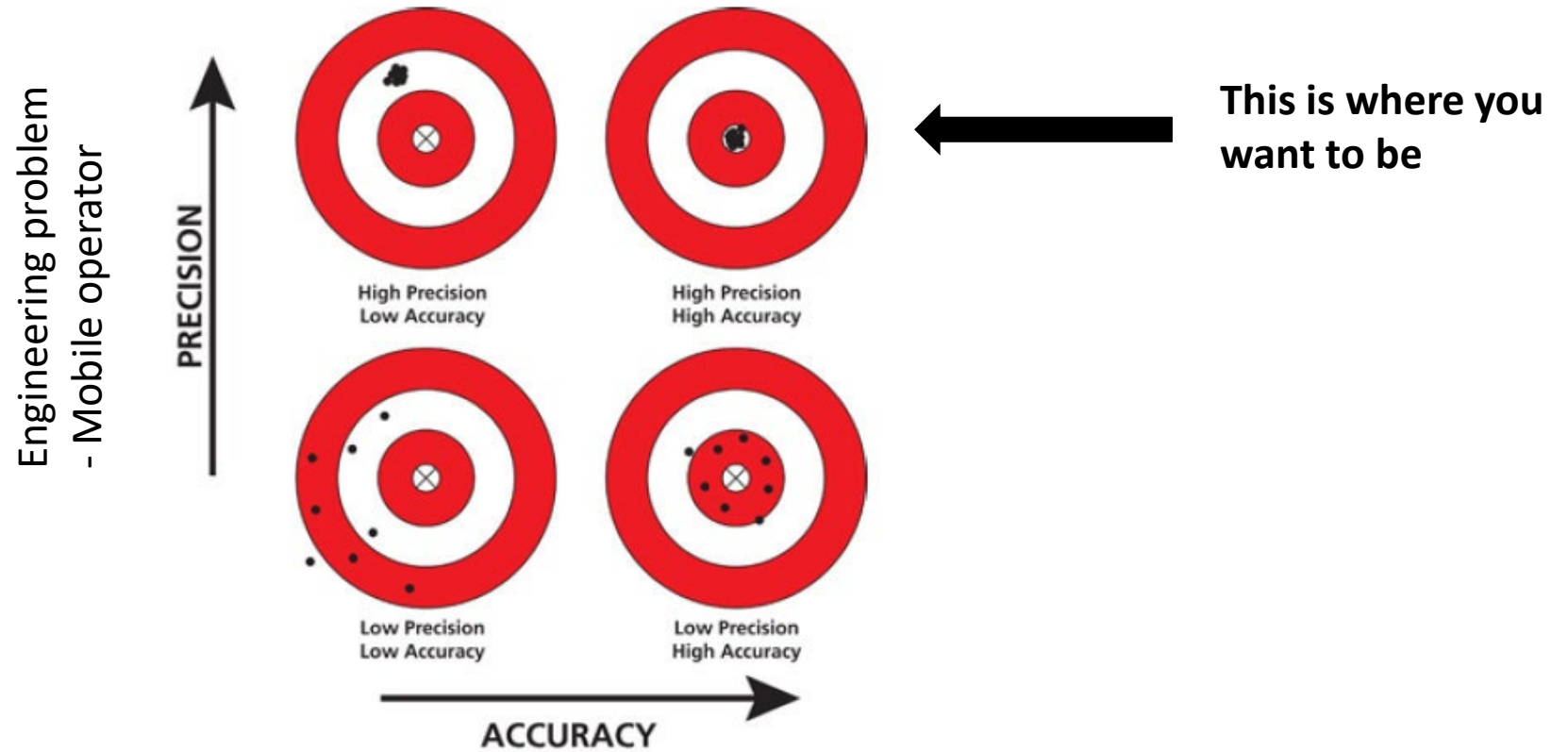


KABUPATEN	RATIO	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
		CDR								
Sanggau	1.99	11,122	16,981	13,770	13,776	17,534	18,642	17,869	10,446	11,451
Natuna	6.01	1,461	1,420	1,537	1,416	1,564	1,636	1,791	2,092	2,113
Malaka	158.75	3	9	50	36	60	54	3	72	31
Kepulauan Anambas	10.78	152	217	176	197	204	193	206	212	481
Kupang	36.64	25	56	137	152	135	99	28	165	52
Rokan Hilir	2.68	1,207	1,271	1,259	1,133	987	987	1,540	542	629
Bengkayang	1.64	1,138	1,972	1,617	1,558	1,602	1,764	2,349	1,211	1,277
Indragiri Hilir	1.76	1,138	1,220	1,271	1,111	1,004	1,245	2,058	779	955
Kepulauan Talaud	25.82	32	63	28	48	55	31	17	44	43
Kapuas Hulu	1.68	921	1,012	1,129	1,001	1,307	1,780	1,862	1,141	917
Pelalawan	1.83	647	619	658	721	758	673	833	632	880
Kepulauan Sangihe	52.57	3	3	13	5	46	1	-	17	9
Sintang	1.93	200	225	245	228	210	246	371	251	282
Keerom	40.42	1	2	8	11	1	7	4	2	2
Malinau	2.63	176	105	95	110	114	129	158	109	125
Lingga	2.30	171	299	297	278	219	196	392	252	196
Boven Digoel	30.23	-	1	9	6	7	5	16	19	31
Pegunungan Bintang	24.00	-	1	-	2	-	1	-	1	1
Mahakam Ulu	21.20	3	1	-	-	1	-	1	4	1
Grand Total	3.47	18,400	25,477	22,299	21,789	25,808	27,689	29,498	17,991	19,476

Signalling contributes most in hard-to-reach areas



The Recipe for Quality Data



Data science and business problem
- Data scientists and statisticians

Level of recommended **spatial** accuracy

	Inbound	Domestic
Estonia = CDR →	County	Local
Indonesia = Signalling	Local	Event
Georgia = CDR	County	Local

Level of recommended temporal accuracy

	Inbound	Domestic
Estonia = CDR →	Monthly	Daily
Indonesia = Signalling	Daily	Hourly
Georgia = CDR	Monthly	Daily