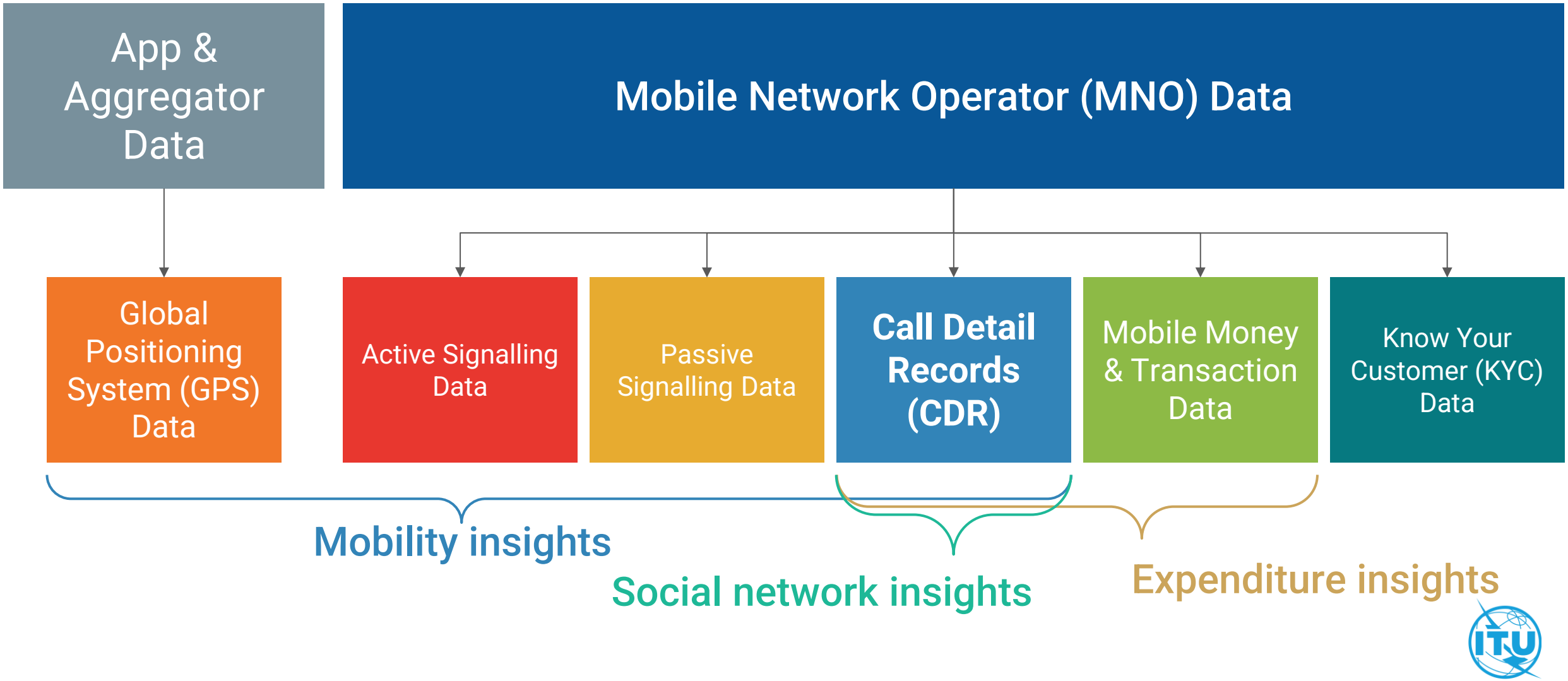
1. What is mobile phone data (MPD), use cases and the typical process?

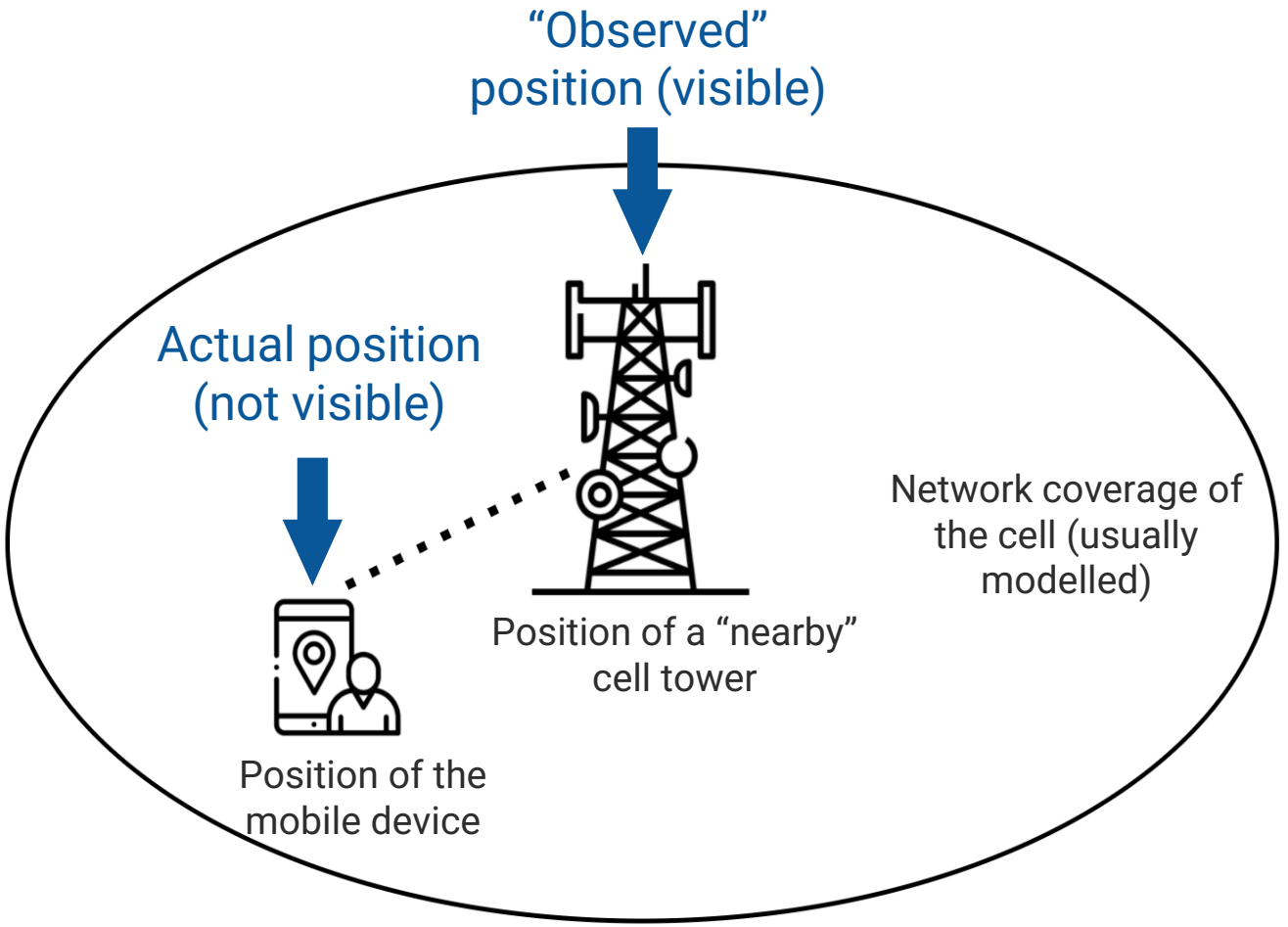
Mobile Phone Data (MPD)



How CDR/IPDR data is generated

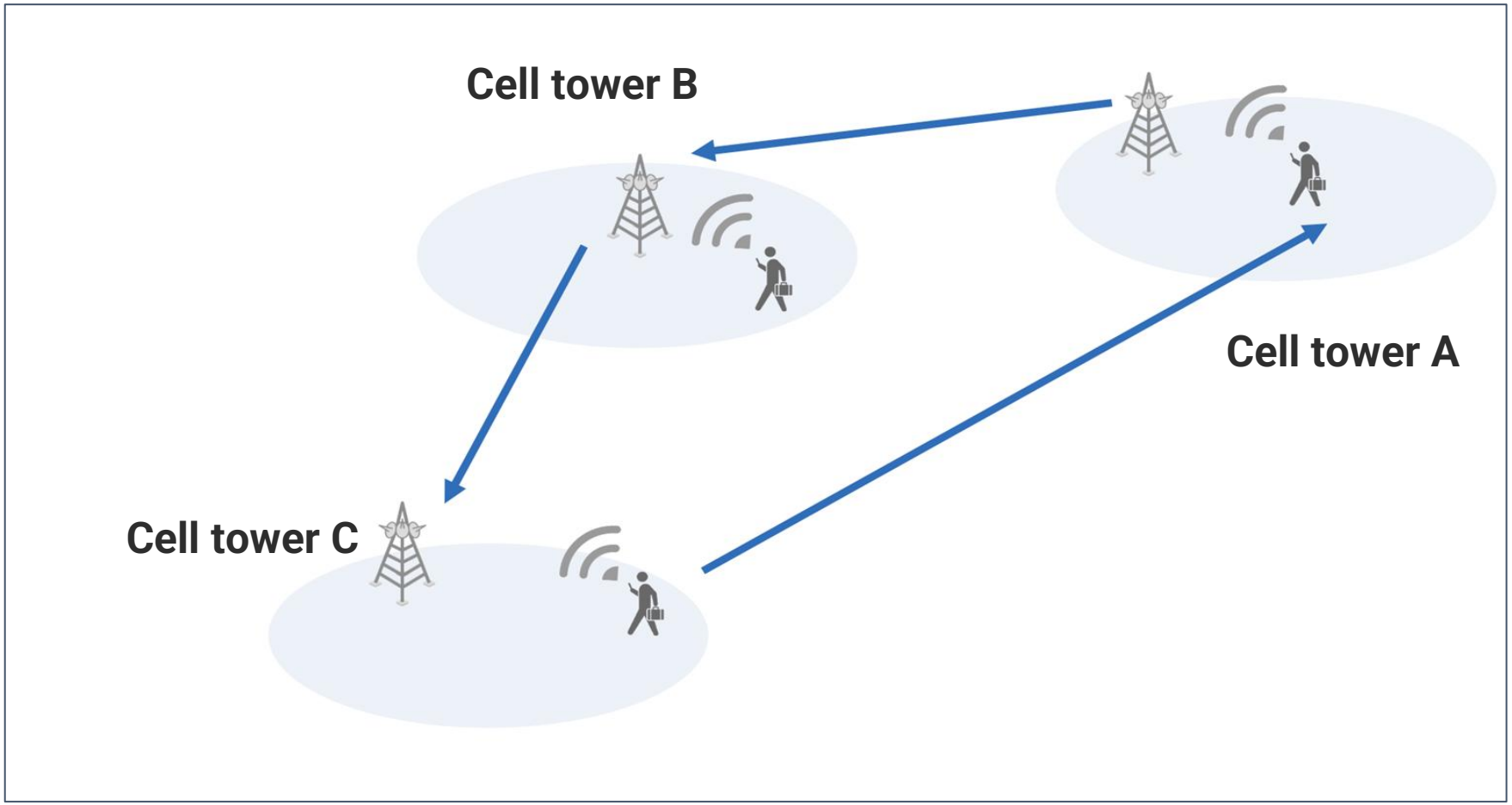
CDR data

- Passively generated when a subscriber
  - Makes or receives **a call**
  - Sends or receives **an SMS**
  - Uses mobile **data**
- Location is at **the cell tower** level
- Routinely stored by MNOs for billing purposes



Device location is observed as cell tower location

Time	Cell Tower
6.00	A
6.35	A
7.25	B
8.00	B
9.40	C
10.45	C
13.00	C
14.50	C
16.00	C
17.30	C
19.00	C
20.20	A
20:50	A
21.35	A
22.50	A
23.45	A



Statistics areas where MPD can be used

1) Tourism statistics

2) Migration statistics

3) Census and dynamic population

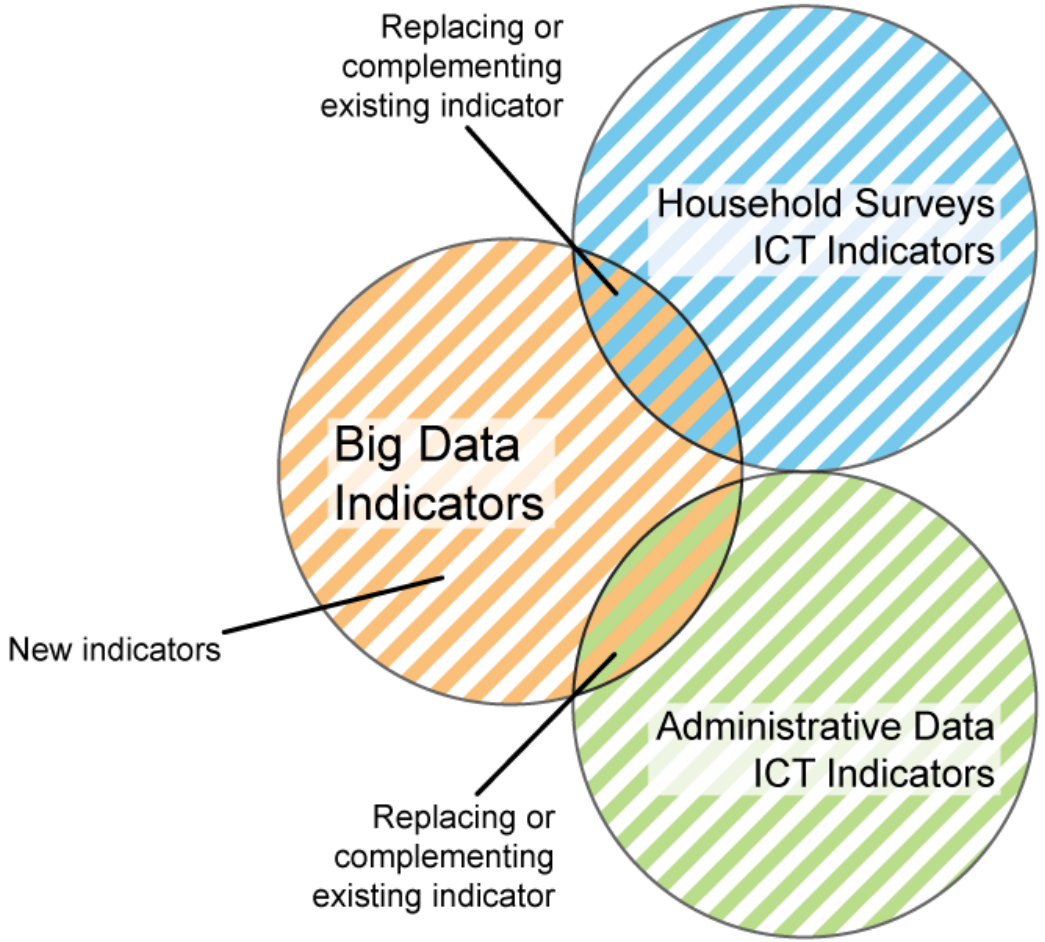
4) Displacement in disaster

5) Information society indicators

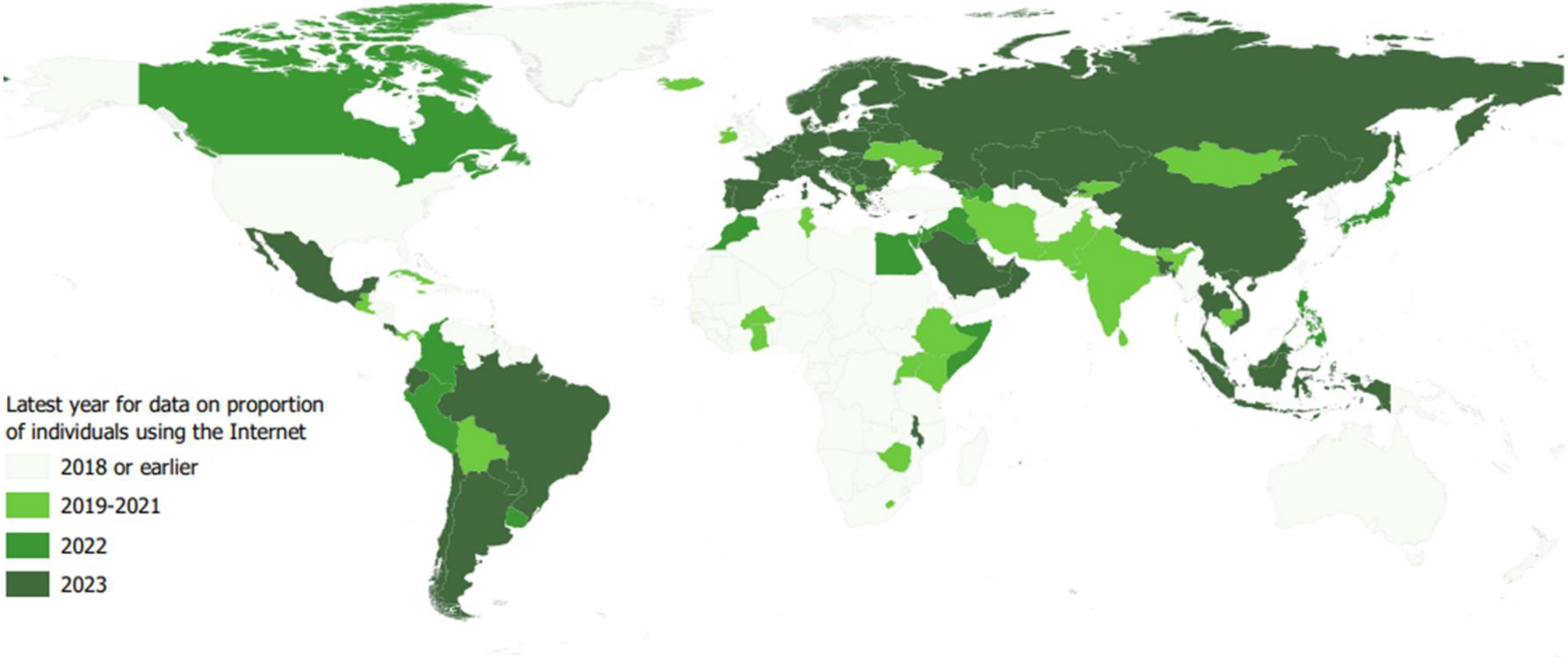
6) Transport and commuting statistics

Big data indicators complementing official information society statistics

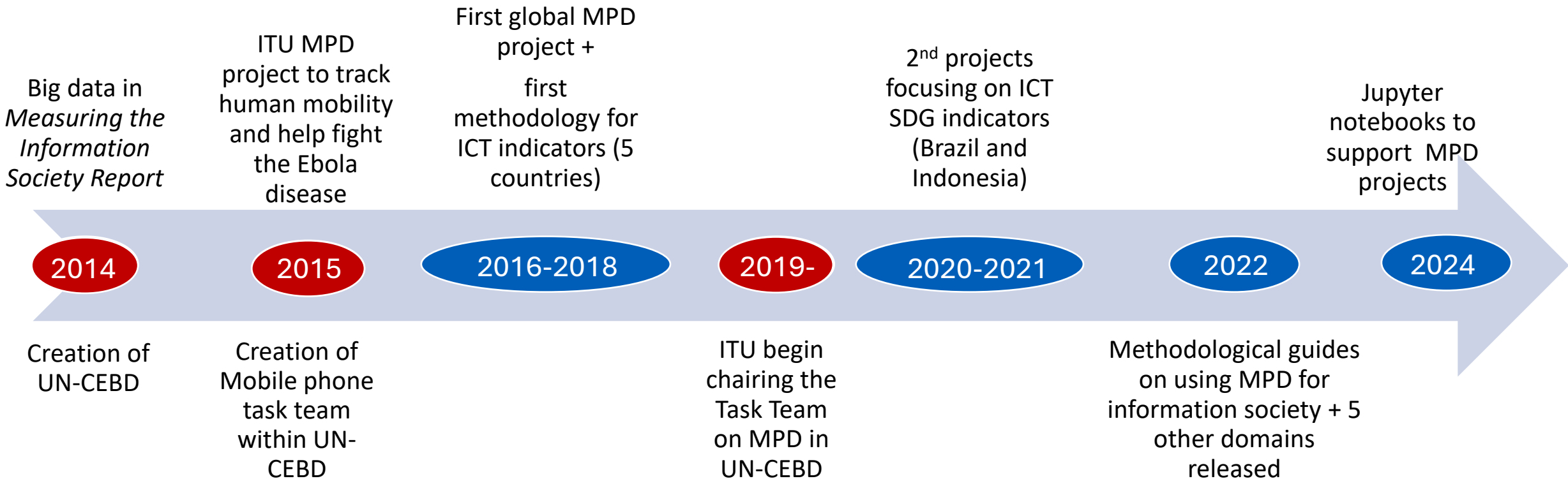
- Improve estimates of key indicators
- Provide more geographic granularity
- Provide new indicators
- Alternative method for comparison



Why looking at MPD for Internet use data?



More than 10 years of using mobile phone data at ITU



ITU MPD for Information Society

1st pilot: 2016-2018

✓6 countries (Colombia, Georgia, Kenya, Philippines, UAE)

✓16 ICT indicators (administrative data)

2nd pilot: 2020-2021

✓ Brazil, Indonesia

- ✓ **2 SDG ICT indicators**
  - ✓ 9.c.1 – Percentage of population covered by mobile network: 2G, 3G and 4G and above (administrative data)
  - ✓ **17.8.1 – Percentage of population using the Internet (household survey data)**

Ongoing: 2023-2025

- ITU/WB GDF-MPD project: 18 countries
- SADC countries
- Malaysia, Azerbaijan, Liberia, DR Congo, Dominican Rep., Maldives, Uganda, Uruguay, etc

Stakeholders and their roles

1. *Telecommunication Regulator / ICT ministry*
   - ✓ Request mobile phone data as a requisite for monitoring licensing condition
   - ✓ Invested in equipment and expertise to store and process these records
   - ✓ Frequently interact with operators in the course of their regulatory work
   - ✓ Positioned to negotiate and mediate access to mobile phone data
2. *National Statistics Office*
   - ✓ Statistical Act
   - ✓ Mandate to produce official statistics and collect the data
   - ✓ Skills to analyse statistical information
3. *Mobile Phone Operators and service providers*
   - ✓ Custodian of mobile phone data
   - ✓ Invested in equipment and expertise to store and process these records
   - ✓ Required to submit records to a regulatory agency as a condition of their licence or franchise
   - ✓ Have staff that have big data skills to analyse MPD
4. *Data Protection Authority*
   - ✓ Provides guidance and oversight for lawful data processing
   - ✓ Ensures safeguards are in place to ensure privacy (pseudonymisation or anonymisation)

Information society indicator – SDG Indicators 17.8.1

- Goal 17: Strengthen the means of implementation and revitalise the global partnership for sustainable development

- Target 17.8: Enhance the use of enabling technology

- **Indicator 17.8.1: Proportion of individuals using the Internet from any location in the past 3 months**

Methodological Guide on Big Data for measuring the SDG Information society indicators

1. Introduction
2. Background
3. Access and preparations
4. Data sources (description of mobile operator data, quality assurance of raw data)
5. Reference data (local admin units, world population, cell data, digital elevation, household survey data)
6. Data processing (models, data protection guidelines)
7. Calculating the indicators (rationale, definition, indicators calculation, quality assurance)
8. Quality assurance
9. Conclusions

- with experiences and examples from country pilots

What is indicator 17.8.1 (using MPD)?

- Proportion of people using mobile phone data (Internet)
- Breakdown by technology: 2G, 3G, 4G/LTE, 5G, etc
- Breakdown by geography: local administrative units (LAU)
- Where subscribers live (home anchor)
- What is the dominant/highest technology used

Assumptions

- Necessary data (Mobile Phone Data (MPD) and reference data) are available
- Hardware and software tools are available
- Necessary skills are available

Data requirements:

1. Data from mobile network operators (MNOs)
2. Reference data
   - Local Administrative Units (LAU)
   - World Population
   - Cell location data
   - Digital elevation model
   - Household surveys and microdata

A Team of Experts: Diverse skills

The team to work on MPD should normally be composed of these and other staff to make sure to cover all the necessary skills

Sys-op for managing the server, software and file system (where MPD, cells and reference data, etc. are stored)

Data engineers who prepare reference data, raw data files and set up configuration

GIS specialist who conducts QA on reference data (administrative units, roads, etc.)

MPD QA specialist who prepares the MPD files received from MNOs and conducts data QA

Domain methodology experts who consult on configuration files and methodology adjustment

Experienced Project Manager to manage all those activities

Necessary technical skills from one example MPD project that mainly used PostgreSQL:
Linux OS and file system, bash and awk, Python3, SQL, PostgreSQL10+, PostGIS2.5+, Apache2, PHP, RESTful API, understanding data security principles and secure data transfer methods

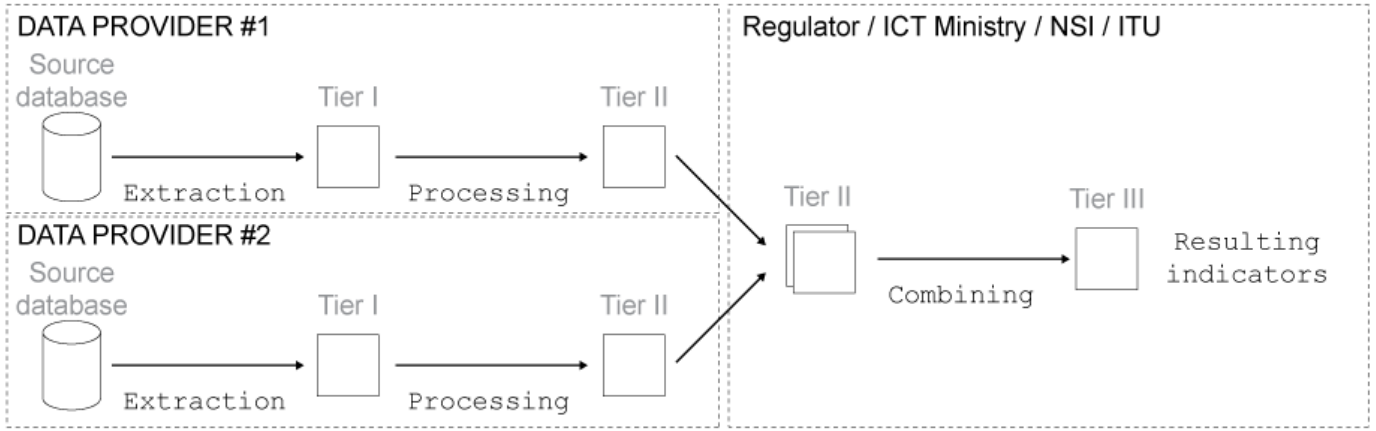
Deep, specific

Level of technical expertise

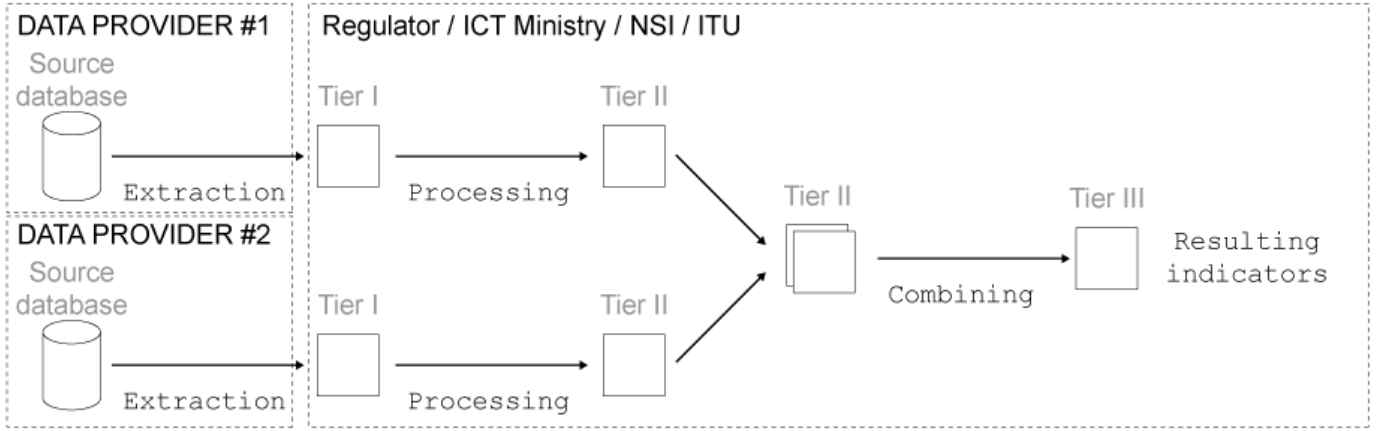
General, overview

Data access models

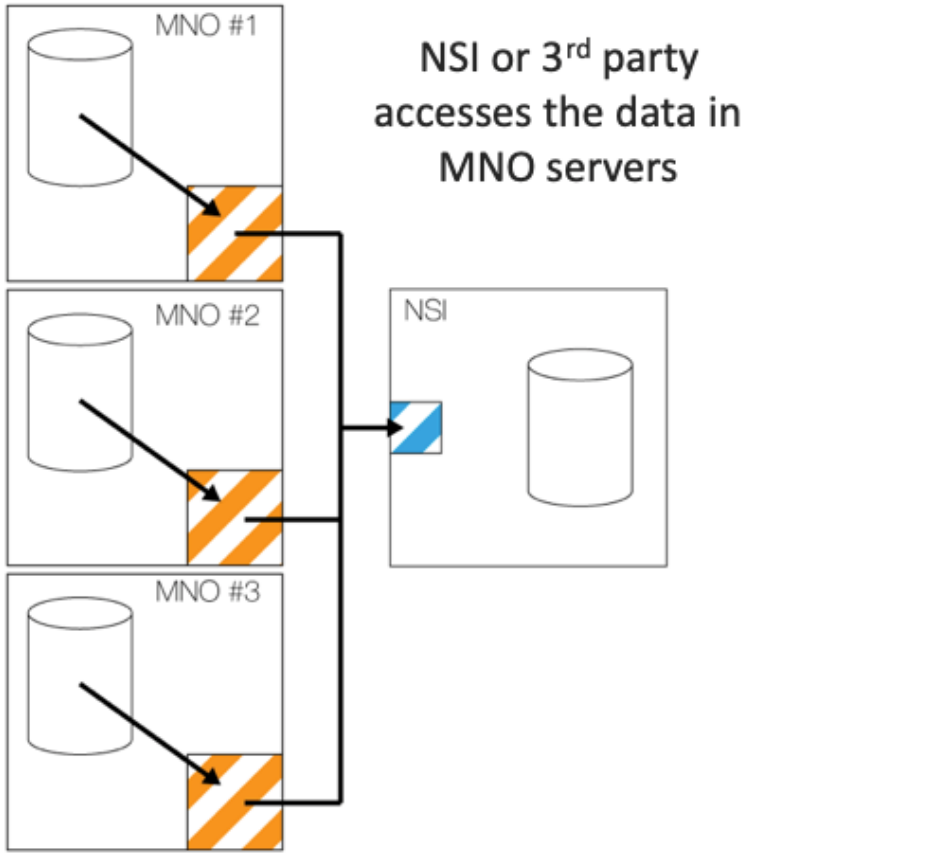
1. Operator-led



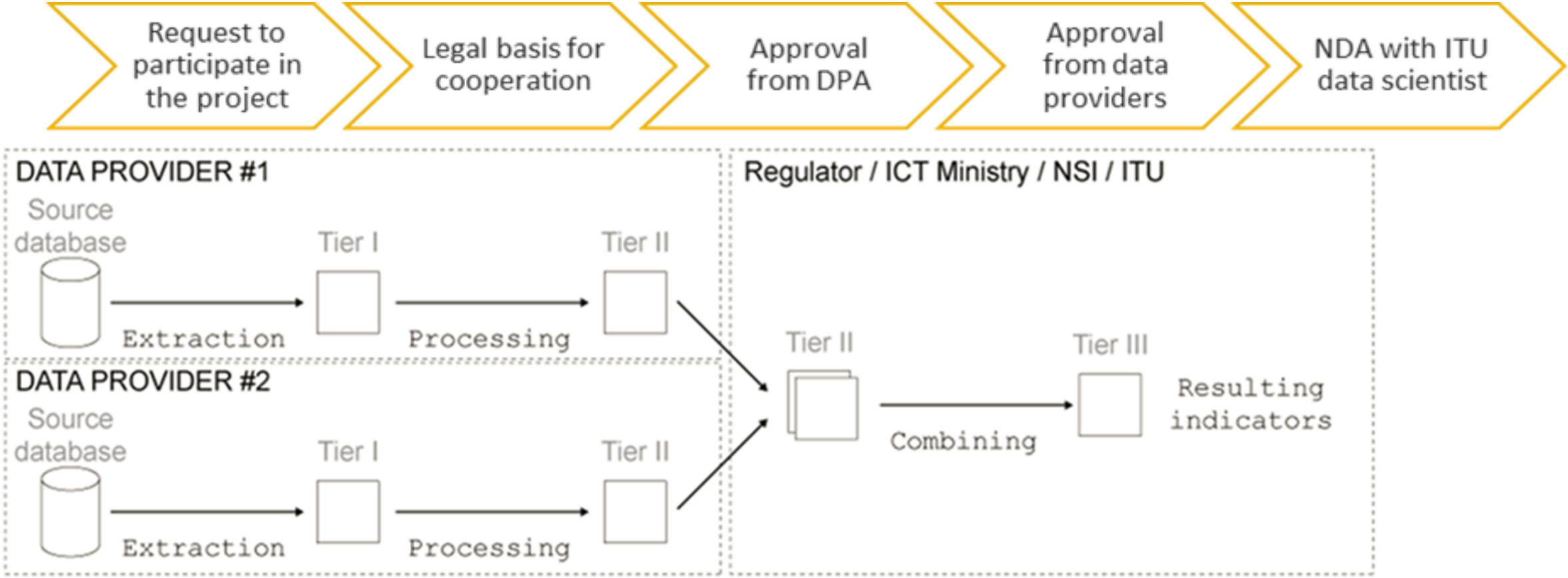
2. Agency-led



3. Public-private partnership

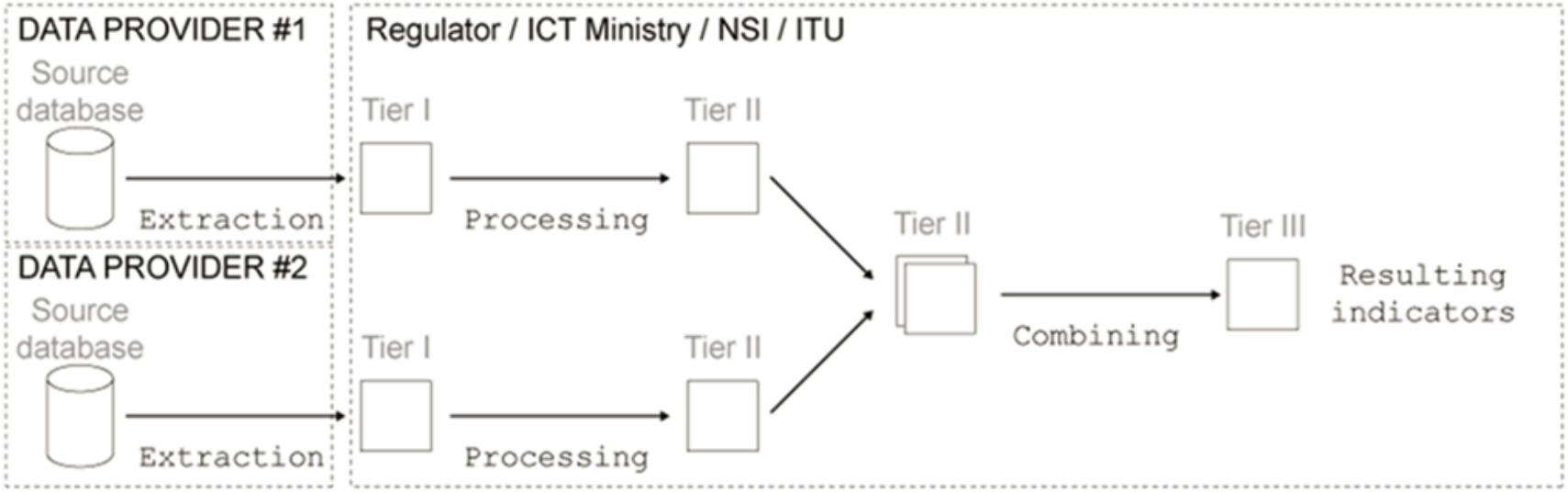


Data Access Model (1)



- Smart, Globe Telecom (MNOs) ----------→**DICT (Philippines)**
- Data for 3 months (April to June 2016)

Data access models (2)



Rio de Janeiro Metropolitan Region

- Two months (2019, March and April)
- One operator (out of four) → market share ≈ 40%

→ IBGE, Brazil

Process of analyzing MPD for ICT indicators

1. Data quality checks of raw data – Sanity/preprocessing
2. Processing of raw data (centralized, distributed, PPP)
   Ensuring privacy and data protection
3. Establishing the home location
4. Calculation of indicators
5. Visualizations
6. Quality assurance of calculated data

2. Requirements (including synthetic data)

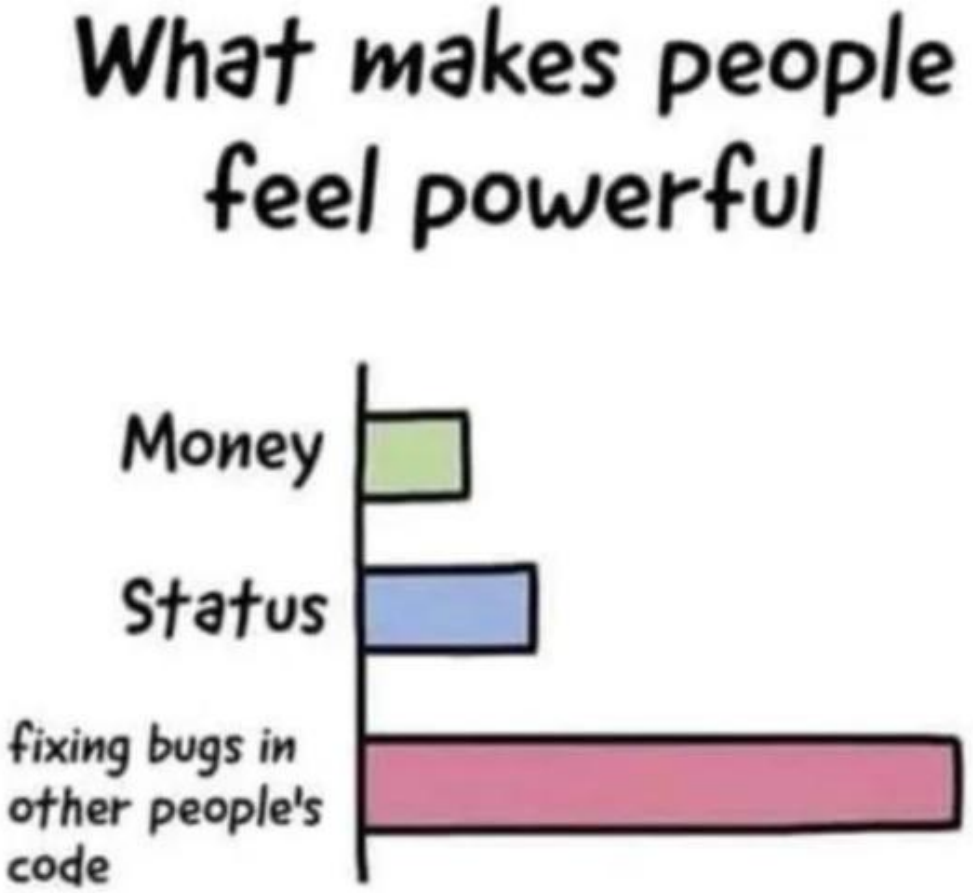
What's the purpose of these codes?

Aim: practical introduction to working with mobile phone data

- Series of Jupyter notebooks "from a to z"
- Codes written in PySpark (Python API for Spark)
- Documentation included in the notebooks

Requirements

- Environment to run Jupyter Notebooks (local, cloud)
- PySpark and Python packages and dependencies installed (Setup code available)
- Raw / synthetic mobile phone data in CSV / Parquet
- Configurations set in config file.
- Staff/skill to adapt code to local circumstances



Data Structure

Mobile network operators (MNOs) collect CDRs for billing purposes and store it when phones connect to nearby cell towers.

Field Name	Type	Mode	Description
msisdn	String		Hashed subscribers identifier
datetime	Timestamp		Transaction date (date and hour)
cell_id	String	NULLABLE	Hashed cell identifier
latitude	Float		Latitude of Base Transceiver Station (BTS)
longitude	Float		Longitude of Base Transceiver Station (BTS)
data_type	String		Data source, can be CDR/CHG or IPDR/UPCC
service	String		Transaction service (4G/ 3G/ 2G)

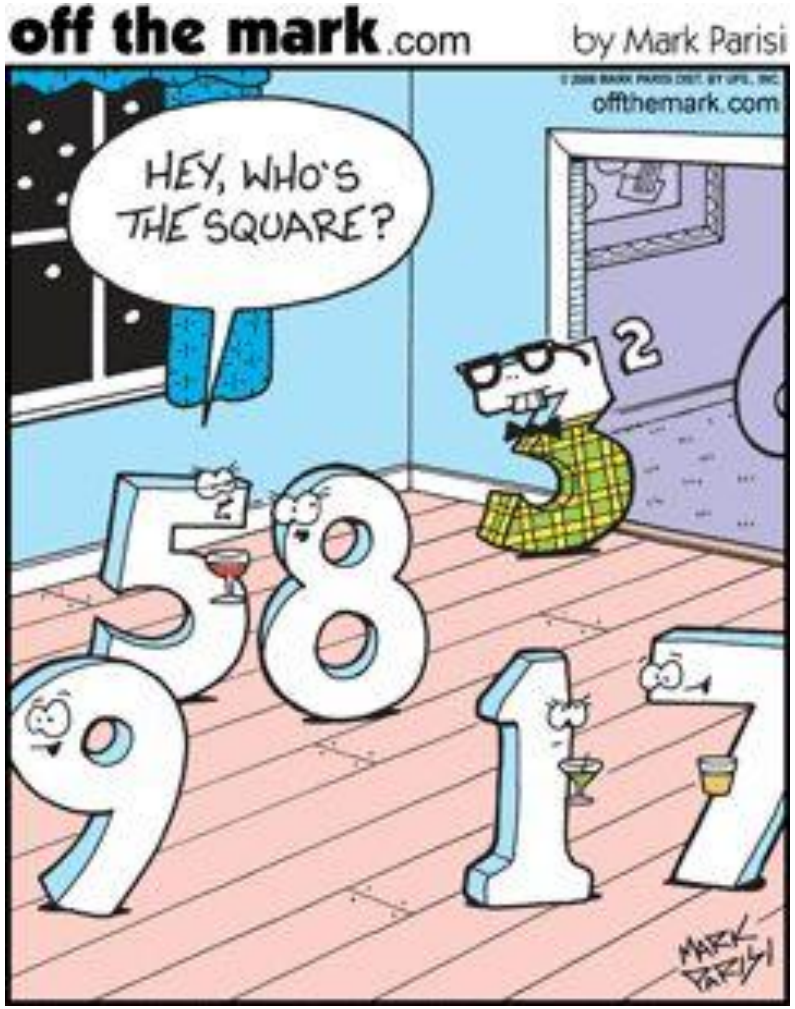
Minimum required fields.

Field name	Type	Mode	Description
msisdn	String		Hashed subscribers identifier
age	Int		Subscribers age from registration data
gender	String		Subscribers gender (M/F) from registration data

Nice to have. Useful for analysis

No data yet? Using synthetic data for training

Artificial data that mimics the statistical patterns and properties of real-life data.



Synthetic MPD data generation

- Synthetic data offers a valuable solution for code/ML development by creating realistic and privacy-preserving datasets.
- Various methods exists for synthetic data generation
- Our scope limited to training and methodology development
- Rules-based synthetic data generation within constrained environment with parameter settings

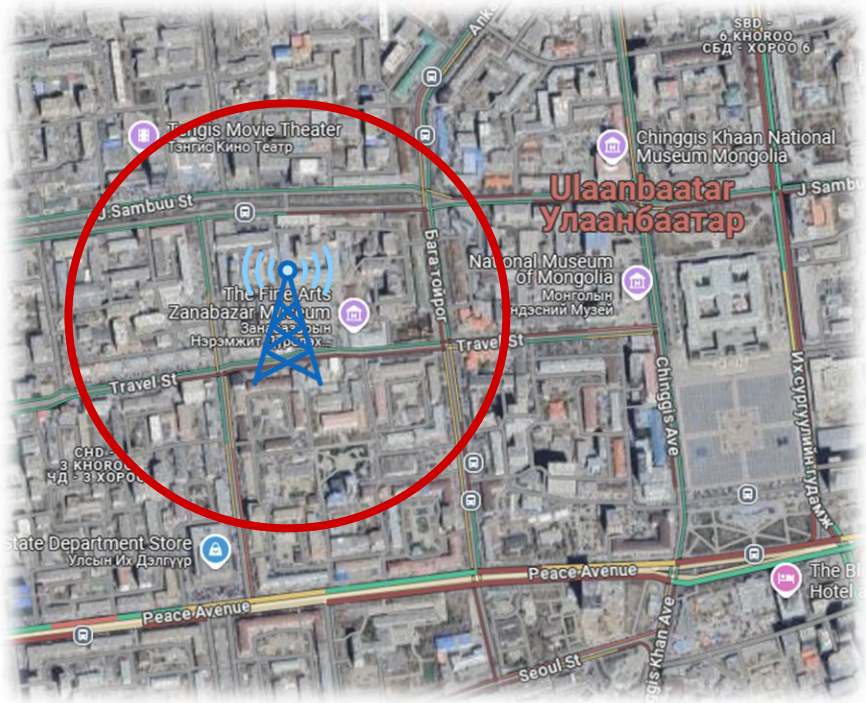
Reference data needed

- Administrative borders in geojson format
  (Recommendation: GADM available by country)
- Cell location data
  (OpenCellID was used for demo / synthetic data)
- Population
  (WorldPop 100x100 m was used for demo / synthetic data)

Base layer prepared using cell location and population to more realistically generate subscriber activity

Notebook 1: Creating a synthetic base layer

- Collect fixed cell locations *
- Calculate buffer zone around each cell (≈ 500m)
- Collect population data
- Map population living within the buffer zone

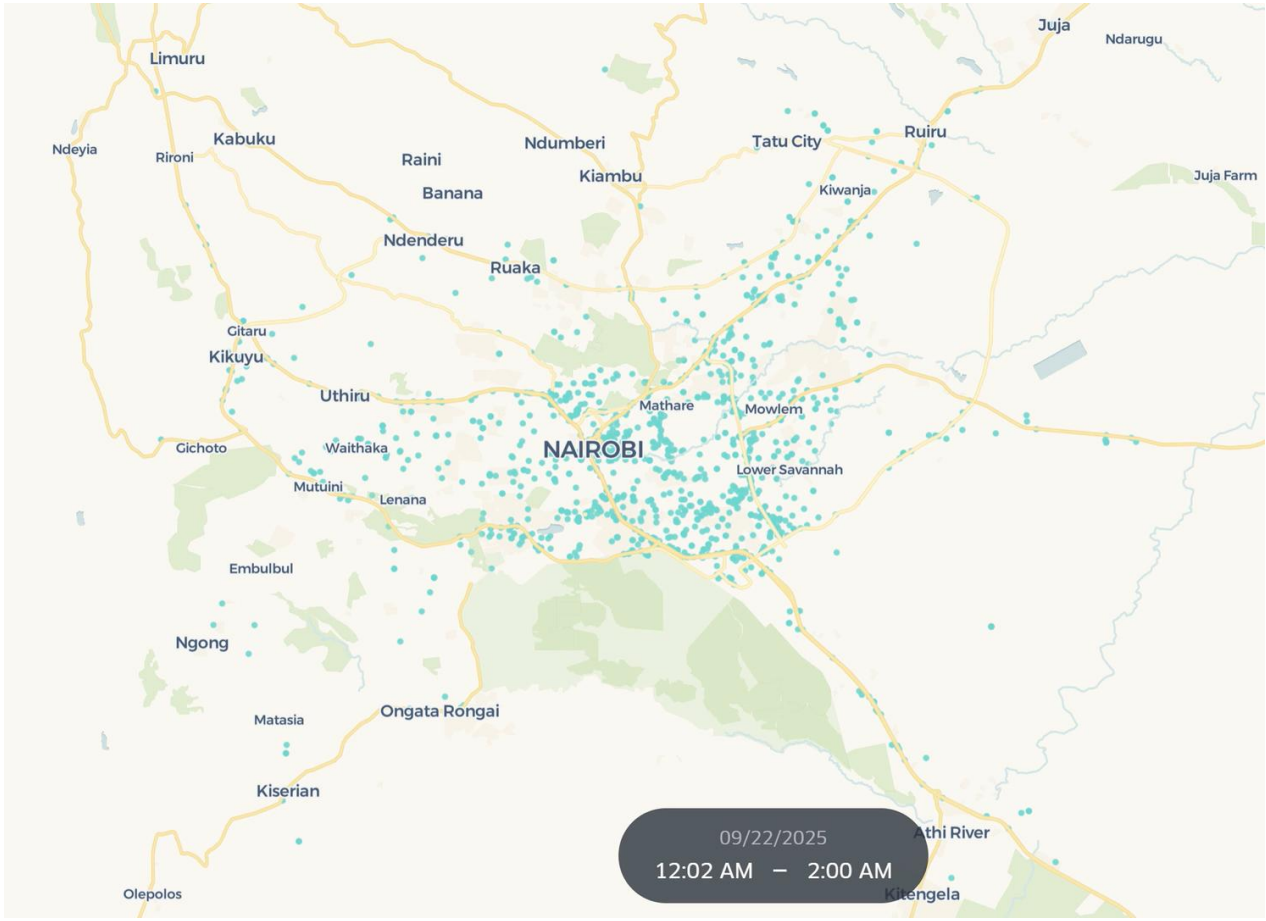


* The notebooks uses Worldpop for granular population estimates and OpencellID for cell locations, if actual cell locations are not available and can thus be run for any part of the world.

radio	cell_id	lon	lat	population	countryid	country	country_lon	iso	network	extracted_date
GSM	0	106.7847	47.9139	6799.4937	152	Mongolia	Mongolia	MNG	Mobicom	2/16/2025
LTE	3	106.6654	47.8985	0	152	Mongolia	Mongolia	MNG	Mobicom	2/16/2025
LTE	5	106.8514	47.9126	7689.266	152	Mongolia	Mongolia	MNG	Mobicom	2/16/2025
LTE	6	106.8906	47.9229	12421.404	152	Mongolia	Mongolia	MNG	Unitel	2/16/2025
LTE	50	106.9149	47.9402	5373.119	152	Mongolia	Mongolia	MNG	Unitel	2/16/2025
UMTS	84	106.6738	47.8992	0	152	Mongolia	Mongolia	MNG	Mobicom	2/16/2025
UMTS	110	106.7941	47.9096	6608.4214	152	Mongolia	Mongolia	MNG	Unitel	2/16/2025

Notebook 2: Create network events

- Population around each cell acts as weighting factor when generating events
- A “home” location is generated for each subscriber -> weights for subsequent events are re-calibrated
- Synthetic CDRs and IPDRs are generated based on rules of typical human mobility patterns

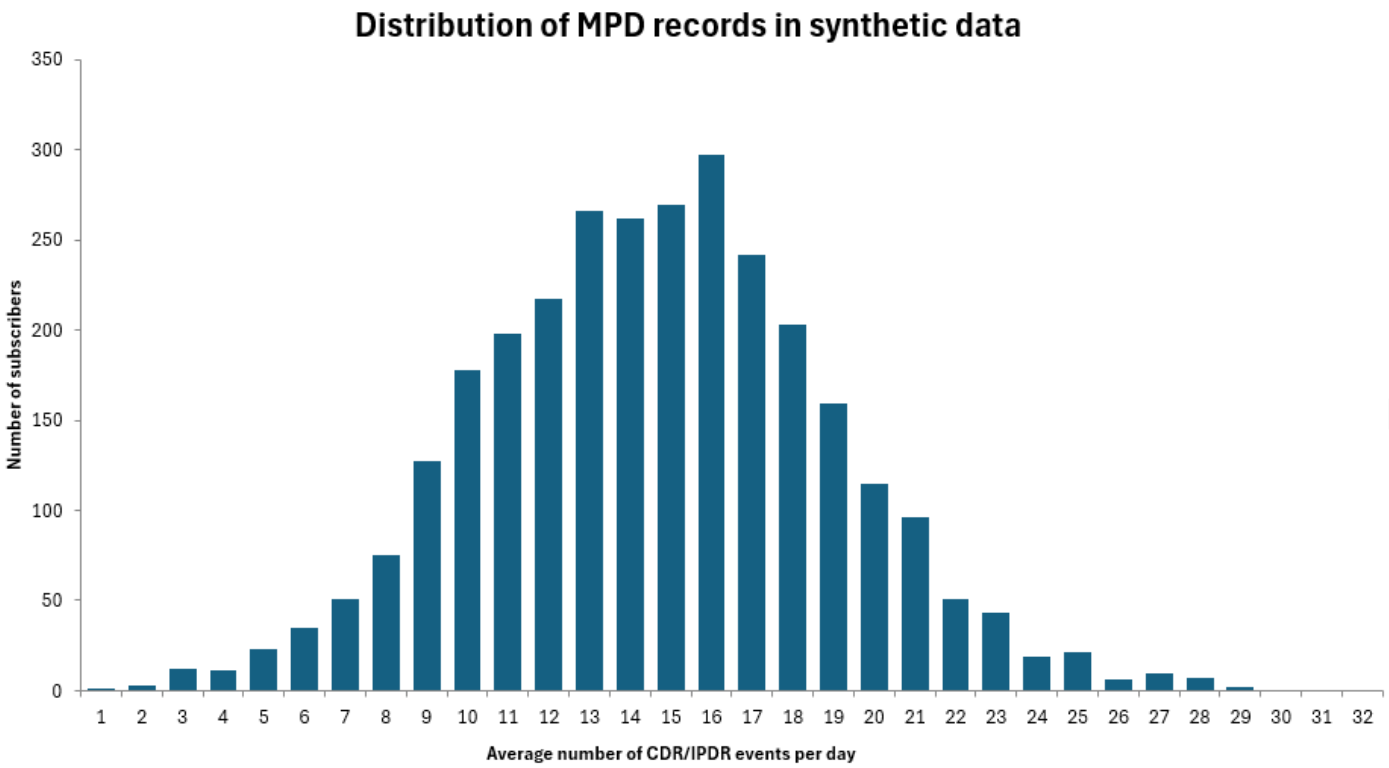


Synthetic CDRs of 3000 subscribers during week of the workshop: (22 Sep – 28 Sep 2025)

Colors:
Dark blue = Weekday, Light blue = Weekday night
Dark red = Weekend, Light red = Weekend night

Parameters and probabilities to create different subscriber profiles

- Number of users
- Time period
- Number of events per user per day
- Distribution of type of events (CDR / IPDR) by technology
- Peak/off-peak hours of events during the day
- Typical area of movement, e.g. 20 km^2
- Typical time during day spent at home and at work
- Weekday / weekend patterns on distribution and location of events
- “Off-time”-probabilities, e.g. sleeping patterns,
- Inclusion of network tests, i.e. “robots”

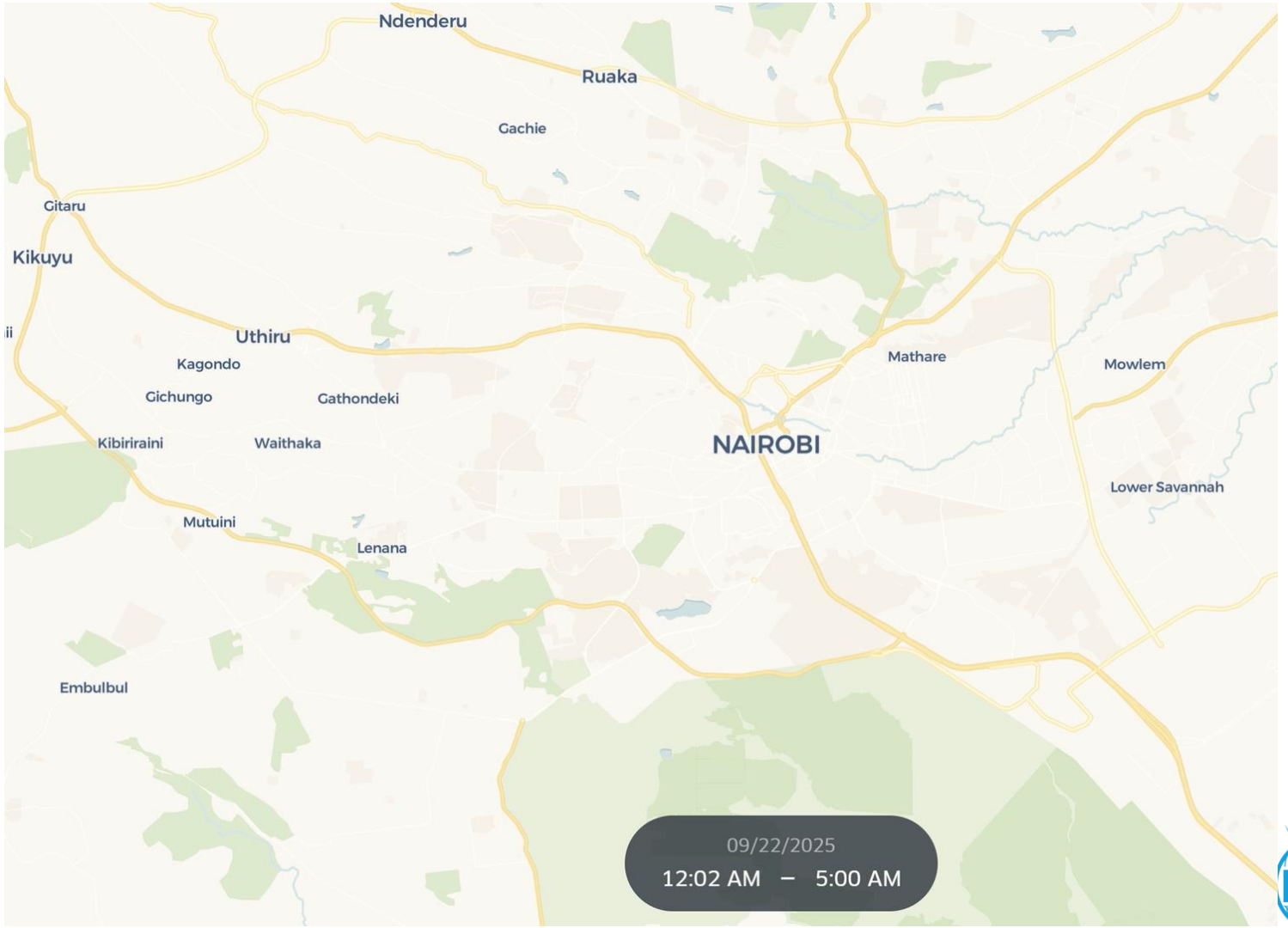


3 000 subscribers = 300 000 records for one week

Example: One subscriber during one week

(22 Sep – 28 Sep 2025)

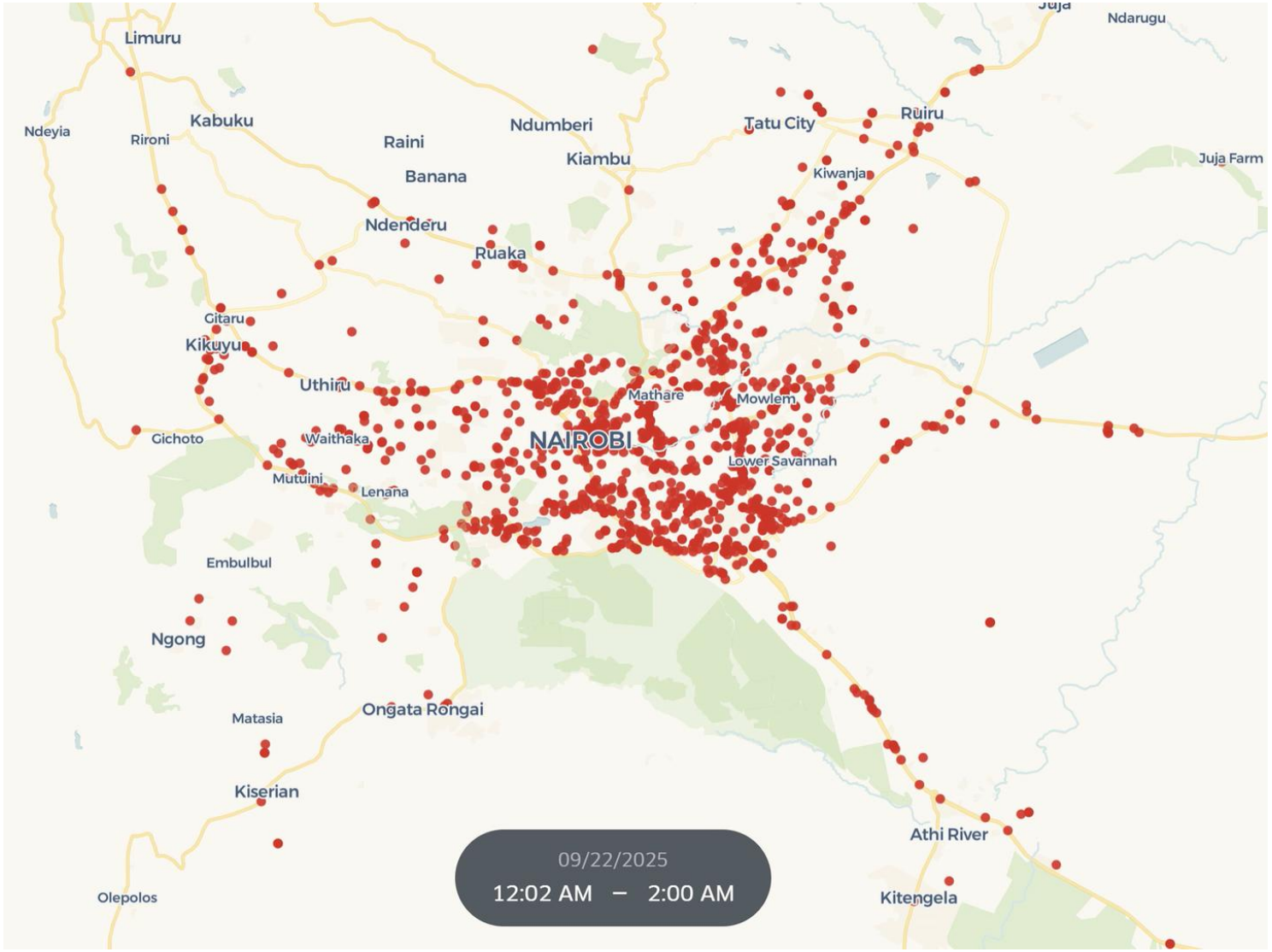
Colors:
Dark red = Home
Dark blue = Work
Green = Common location 1
Purple = Common location 2
Orange = Other location



Example: All subscriber during one week

(22 Sep – 28 Sep 2025)

Colors:
Dark red = Home
Dark blue = Work
Green = Common location 1
Purple = Common location 2
Orange = Other location



1. Sanity checks

Goal: check, clean and transform the raw MPD into a more usable format for analysis

Pre-step: Check the data

```
mno_ms_id|pos_time|mno_cell_id
34f9834rhj384j9384|??2019-01-01 12:00:01|^M10052
34f9834rhj384j9384|2019-01-01 12:15:13|47461
34f9834rhj384j9384|2019-01-01 18:30:24|10052
34f9834rhj384j9384|2019-01-01 18:30:24|10052
34f9834rhj384j9384|2019-01-01 20:04:12|30461
34f9834rhj384j9384|2019-01-01 20:05:55
34f9834rhj384j9384|2019-01-02 00:30:28|20490
34f9834rhj384j9384|02/01/2019 01:23:00|30461
34f9834rhj384j9384|2019-01-02 01:25:10|30461
34f9834rhj384j9384|2019-01-03 55:00:00|20490
```

- hidden characters
- duplicate rows
- missing delimiters
- inconsistency in data types
- impossible timestamps

- Consistency in file structure.
- Elements that are possible source of issues for automatic processes.

1. Sanity checks: Output file

- Output file includes errors only
- Row identification and description to easily identify the inconsistent rows.
- Possibility with multiple errors on the same line
- Output in Json format

```
msisdn,datetime,cell_id,latitude,longitude,data_type,service
??subscribers_00000,2025-03-17 1:21,279,47.914,106.9173,IPDR,3G
subscribers_00000,2025-03-17 4:13,279,47.914,106.9173,IPDR,3G
subscribers_00000^M,2025-03-17 7:37,267,47.955,106.916,IPDR,3G
subscribers_00000,2025-03-17 9:16,267,47.955,106.916,CDR,3G
subscribers_00000,2025-03-17 9:42,267,47.955,106.916,IPDR,3G
subscribers_00000,2025-03-17 11:06,267,47.955,106.916,CDR,3G
subscribers_00000,3/17/2025 11:10,267,47.955,106.916,IPDR,3G
subscribers_00000,2025-03-17 11:44,267,47.955,106.916,IPDR,3G
?subscribers_00000,2025-03-17 12:26,267,47.955,106.916,CDR,3G
subscribers_00000,2025-03-17 12:31,267,47.955,106.916,IPDR,3G
subscribers_00000,45734.48958,267,47.955,106.916,IPDR,3G
subscribers_00000,2025-03-17 14:25,267,47.955,106.916,IPDR,3G
subscribers_00000,2025-03-17 14:30,267,47.955,106.916,IPDR,3G
subscribers_00000,2025-03-17 15:51,267,47.955,106.916,IPDR,3G
subscribers_00000,2025-03-17 18:36,267,47.955,106.916,CDR,3G
subscribers_00000,2025-03-17 21:12,267,47.955,106.916,IPDR,3G
subscribers_00000,2025-03-17 21:47,279,47.914,106.9173,CDR,3G
subscribers_00000,2025-03-17 23:37,275,47.9146,106.9171,CDR,3G
??subscribers_00000,2025-03-18 0:14,279,47.914,106.9173,CDR,3G
```


```
1 {"row": 1, "case_type": [{"case": "Case 2 - Hidden Characters: msisdn", "value": "??subscribers_00000"}]}
2 {"row": 3, "case_type": [{"case": "Case 2 - Hidden Characters: msisdn", "value": "subscribers_00000^M"}]}
3 {"row": 7, "case_type": [{"case": "Case 3 - Inconsistent Data Type: datetime", "value": "3/17/2025 11:10"}, {"case": "Case 4 - Impossible Timestamp", "value":
  "3/17/2025 11:10"}]}
4 {"row": 9, "case_type": [{"case": "Case 2 - Hidden Characters: msisdn", "value": "?subscribers_00000"}]}
5 {"row": 11, "case_type": [{"case": "Case 3 - Inconsistent Data Type: datetime", "value": "45734.48958"}, {"case": "Case 4 - Impossible Timestamp", "value":
  "45734.48958"}]}
6 {"row": 19, "case_type": [{"case": "Case 2 - Hidden Characters: msisdn", "value": "??subscribers_00000"}]}
```

Step 2: Preprocessing – a) drop duplicate rows

- Check and remove any duplicates
- Duplicates can depend on the detail of information, e.g. number of decimals used, e.g, lon-lat or time

```
# Print the number of records in the DataFrame
print("Number of records before deduplication: {}\n".format(df.count()))

# Drops the duplicate rows from the dataframe
df = df.dropDuplicates()

# Display the first five rows of the DataFrame in a tabular format
df.show(5)

# Print the number of records in the DataFrame
print("Number of records after deduplication: {}".format(df.count()))
```

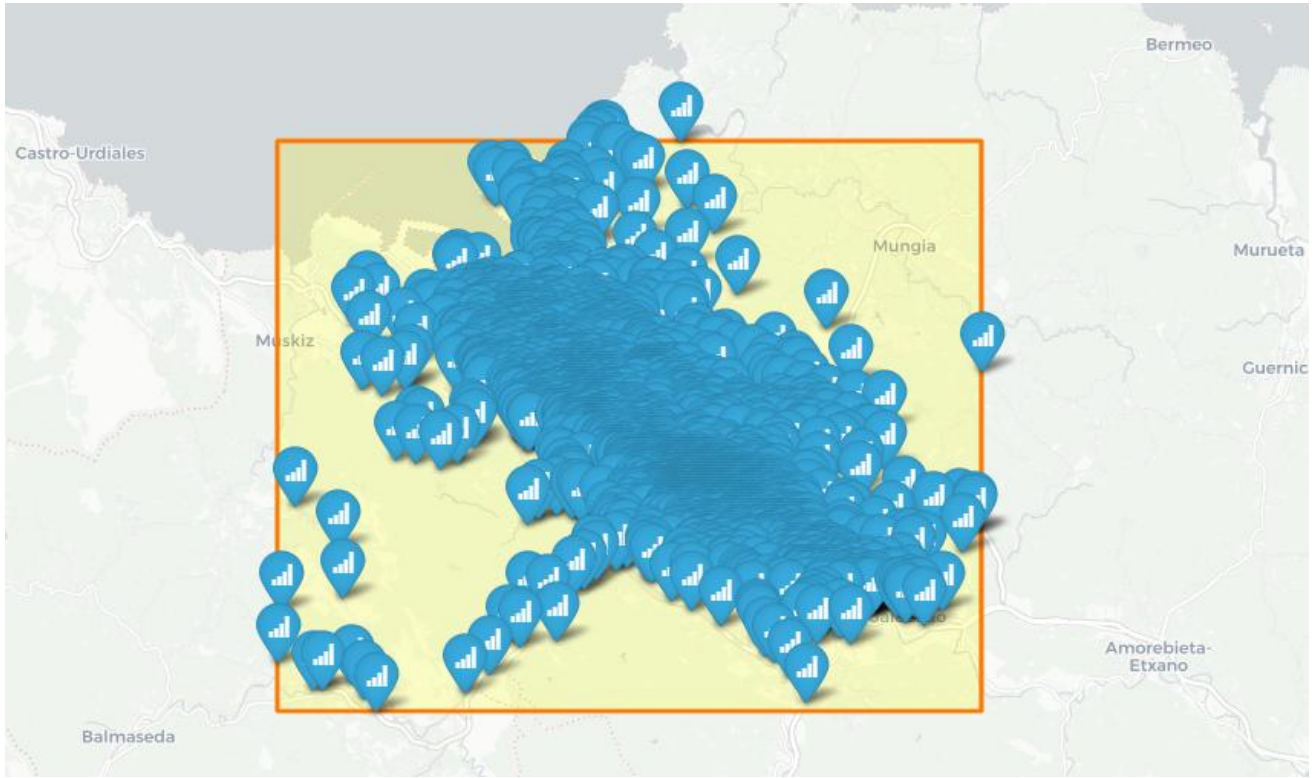
```
Number of records before deduplication: 58427
```

```
+----------------+-------------------+-------+--------+---------+---------+-------+----------+
|          msisdn|           datetime|cell_id|latitude|longitude|data_type|service|      date|
+----------------+-------------------+-------+--------+---------+---------+-------+----------+
|subscribers_00000|2024-06-13 20:34:00|    552|  43.255|   -2.939|      CDR|     2G|2024-06-13|
|subscribers_00000|2024-06-16 05:42:00|    552|  43.255|   -2.939|     IPDR|     2G|2024-06-16|
|subscribers_00001|2024-06-14 13:59:00|   1161|    43.3|   -2.993|     IPDR|     4G|2024-06-14|
|subscribers_00006|2024-06-10 06:49:00|   2855|  43.248|   -2.976|      CDR|     3G|2024-06-10|
|subscribers_00006|2024-06-11 12:45:00|   2855|  43.248|   -2.976|      CDR|     3G|2024-06-11|
+----------------+-------------------+-------+--------+---------+---------+-------+----------+
only showing top 5 rows

Number of records after deduplication: 51182
```

Quick Exploratory Data Analysis (EDA)

- Quick check aggregated statistics:
  - Number of subscribers
  - Number of records per subscriber etc.
- Ensure number of cells and their geographic coordinates etc. seem appropriate

unique_subscribers	unique_cell_id	unique_data_type	unique_service
115	2107	2	3



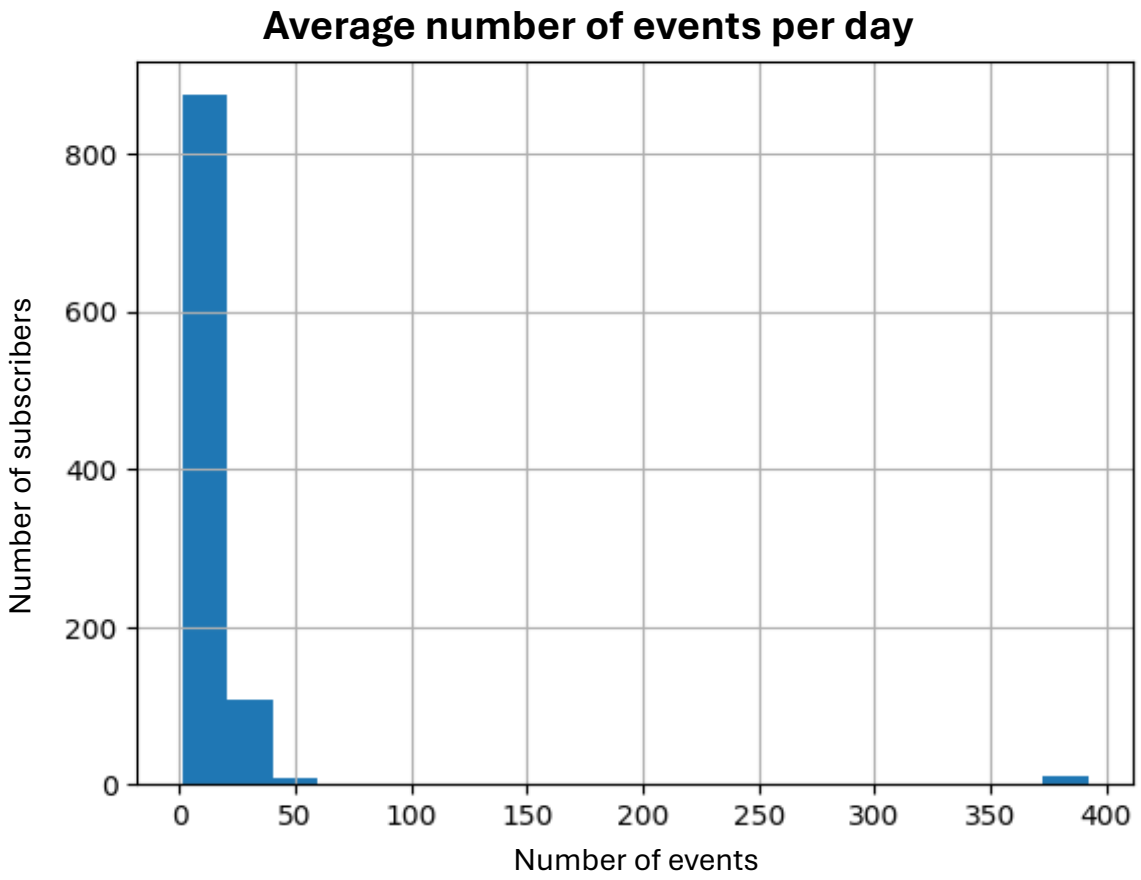
b) “Robot” Filtering

- Outliers with extremely large number of events during a short time period, e.g. tests done by operators
- Code includes a parameter with the number of events considering to be “robots”

```
# Convert the first 10 values of the 'msisdn' column from the `df_robots` DataFrame into a li
df_robots['msisdn'].tolist()[:10]
```

```
Detecting 10 subscribers (8.70%) with more than 200 events in a single site at a single day.
```

```
[20]: ['subscribers_00006',
       'subscribers_00012',
       'subscribers_00017',
       'subscribers_00022',
       'subscribers_00033'
```




```
Number of records before robot filtering: 173481
[Stage 81:=============================================>
Number of records after robot filtering: 147751
```

c) random location / tourist filtering (if needed)

- For some use cases, e.g. information society, not all records are needed.
- Removing *random locations* – if subscriber has one record at a cell location - can reduce the number of records and speed up processing
- *Tourist filtering* removes users with very few records during the entire time period.

```
Number of records before random records filtering: 536310
```

msisdn	datetime	cell_id	latitude	longitude	data_type	service	date	records_site
0	2024-06-21 14:17:00	338.0	43.331	-3.132	IPDR	3G	2024-06-21	2
0	2024-06-21 15:57:00	338.0	43.331	-3.132	IPDR	3G	2024-06-21	2
0	2024-07-17 16:53:00	338.0	43.331	-3.132	IPDR	3G	2024-07-17	1
0	2024-10-02 15:00:00	338.0	43.331	-3.132	IPDR	3G	2024-10-02	1
0	2024-10-13 15:55:00	338.0	43.331	-3.132	IPDR	3G	2024-10-13	1
0	2024-05-10 05:36:00	360.0	43.21	-3.129	IPDR	4G	2024-05-10	1
0	2024-05-13 16:16:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 08:10:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 17:05:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 08:59:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9

only showing top 10 rows

msisdn	datetime	cell_id	latitude	longitude	data_type	service	date	records_site
0	2024-06-21 14:17:00	338.0	43.331	-3.132	IPDR	3G	2024-06-21	2
0	2024-06-21 15:57:00	338.0	43.331	-3.132	IPDR	3G	2024-06-21	2
0	2024-05-13 16:16:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 08:10:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 17:05:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 08:59:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 02:28:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 14:28:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 16:55:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9
0	2024-05-13 09:39:00	369.0	43.218	-3.128	IPDR	3G	2024-05-13	9

only showing top 10 rows

```
[Stage 128:=================================================>  (192 + 3) / 200]
Number of the records after random records filtering: 236580
```

Save pre-processed data

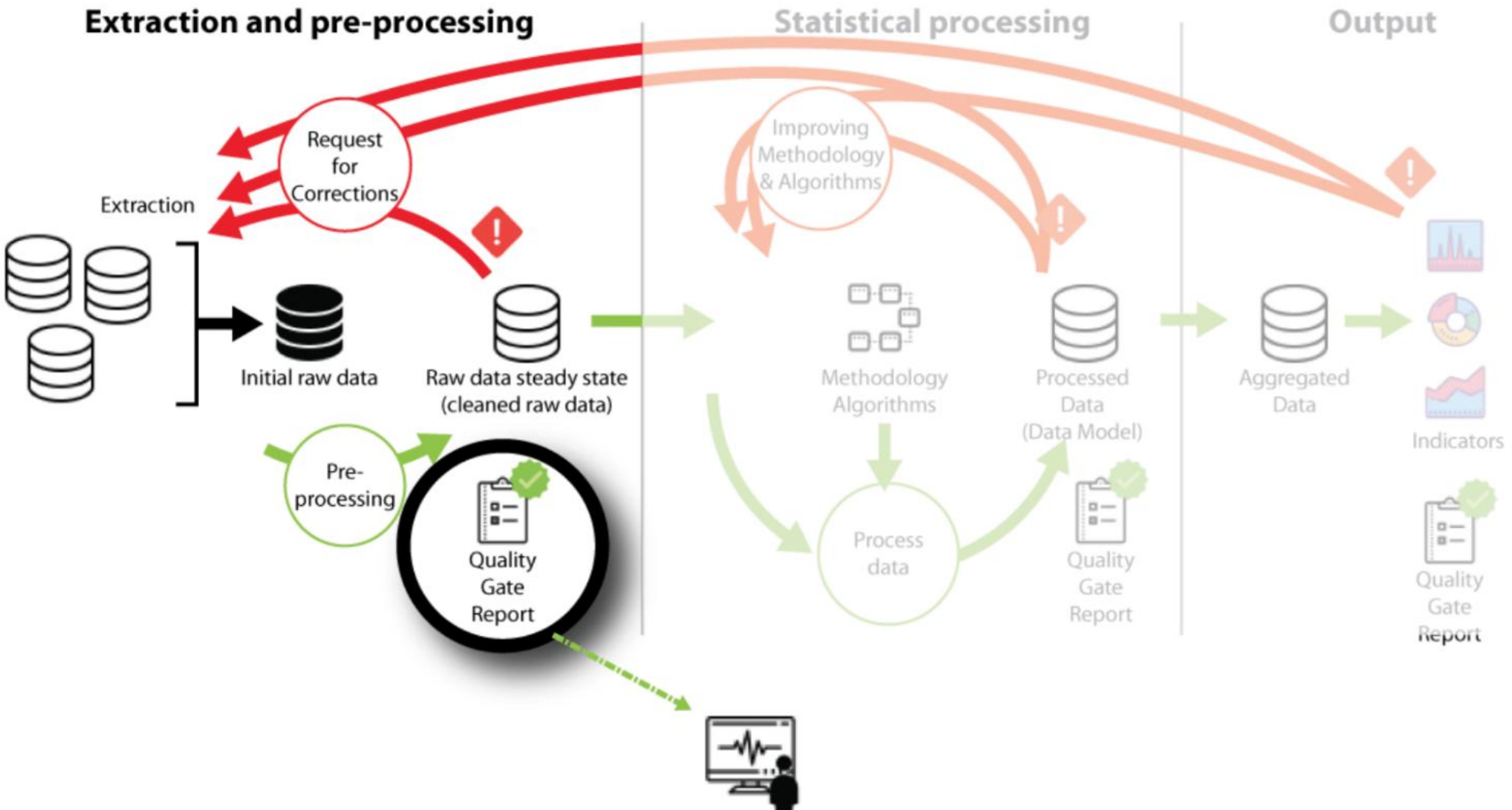
- Filtered data saved to be used in next steps.
- Partitioning often helps with faster processing of large datasets by reading/writing only necessary partition(s) instead of the whole dataset.

```
# Write the DataFrame `df_all_filtered` to disk in the Parquet file format u
# Before write the file, we can drop the records_site and month column since
# The data will be partitioned by the column 'msisdn' using the `partitionBy
# This means that each unique value of 'msisdn' will be stored in a separate
# The output data will be written to the directory specified by 'BASE_PATH',
df_all_filtered = df_all_filtered.drop('records_site').drop('month')

df_all_filtered\
    .write.mode("overwrite")\
    .partitionBy('msisdn')\
    .parquet(BASE_PATH+FILTERED_FILE_PATH_PARQUET)
```

```
df_all_filtered = df_all_filtered.drop('records_site').drop('month')

df_all_filtered\
    .coalesce(1)\
    .write.option("header",True)\
    .mode("overwrite")\
    .csv(BASE_PATH+FILTERED_FILE_PATH_CSV)
```

3. Quality Assurance

3. Demo - Quality Assurance



MPD Quality Assurance (QA)

- Input data QA is important:
  - Garbage in – Garbage out (GIGO)
  - To evaluate the suitability of available raw MPD for the purposes of generating statistics.
  - Describe the dataset and identify aspects that may cause bias and affect coverage, frequency, quality and accuracy.
- QA indicator categories are:
  - Critical
  - Important
  - Nice to have
- The result of each QA indicator can be:
  - Positive (Passed)
  - Acceptable with reservations
  - Needs improvements (not acceptable) (Failed)

MPD Quality Assurance (QA)

1. Generation of user/subscriber statistics
2. Evaluation of the percentage of null values for each
3. Consistency in the number of subscribers per day
4. Consistency in the number of unique cell locations
5. Examination of the location of cell IDs outside the mainland
6. Analysis of the distribution of active days for subscribers (day present)
7. Evaluation of the diurnal distribution of subscribers' activity

Generation of user/subscriber statistics

This metric involves generating comprehensive statistics on the users and subscribers present in the MPD.

```
# Adds the user statistics report to the QA_summary dictionary
QA_summary["USER_STATS_REPORT"] = ("OBSERVED",user_stats_report)
```

```
[Stage 12:=========================================>               (5 + 2) / 7]
Overall start date              : 2024-06-10
Overall end date                : 2024-06-16
Subs. with CDR & IPDR    : 100
Subs. with CDR only             : 100
Subs. with IPDR only     : 100
```

Evaluation of the percentage of null values

```
print(null_report)
QA_summary['NULL_REPORT'] = null_report

# display the entire null_percentages_df dataframe
null_percentages_df
```

Out[9]:

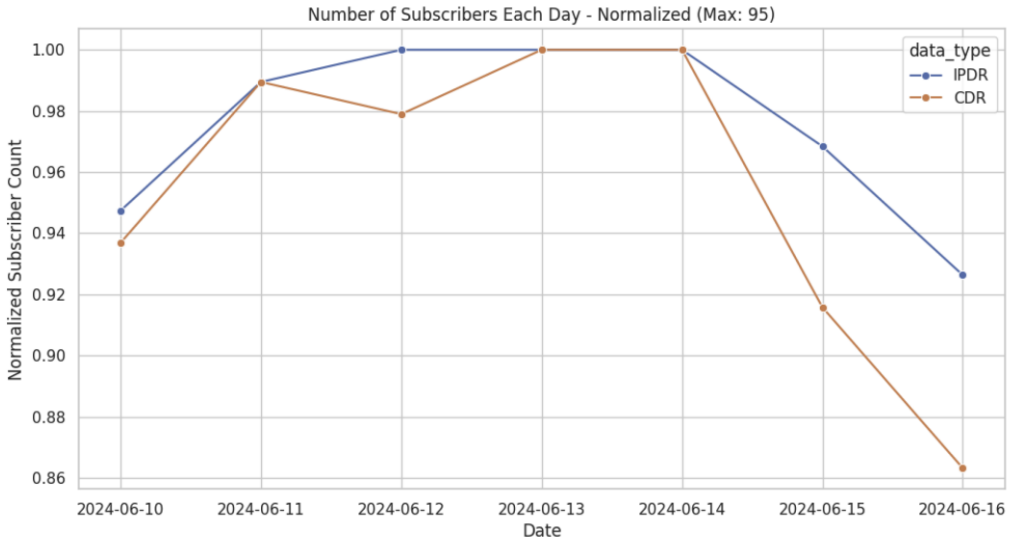
	Column	NullPercentage
0	datetime	0.0
1	cell_id	0.0
2	latitude	0.0
3	longitude	0.0
4	data_type	0.0
5	service	0.0
6	date	0.0
7	msisdn	0.0

Passed: Any field has less than 5% of missing values.

Number of subscribers per day

This metric aims to verify the consistency of the number of subscribers recorded each day in the MPD dataset. Inconsistent subscriber counts may indicate data quality issues, such as data entry errors or inconsistencies in data collection processes.

```python
import numpy as np
from pyspark.sql.functions import countDistinct

# Select the columns for analysis
subscriber_counts_df = df.select("date", "msisdn", "data_type").dropDuplicates()

# Group the dataset by 'date' and 'data_type' then count the number of unique subscribers for each day
subscriber_counts_df = subscriber_counts_df\
    .groupBy('date','data_type')\
    .agg(
        countDistinct('msisdn').alias('subscriber_count')
    )

# Calculate the minimum and maximum subscriber counts
min_count = subscriber_counts_df.agg({'subscriber_count': 'min'}).collect()[0][0]
max_count = subscriber_counts_df.agg({'subscriber_count': 'max'}).collect()[0][0]
```



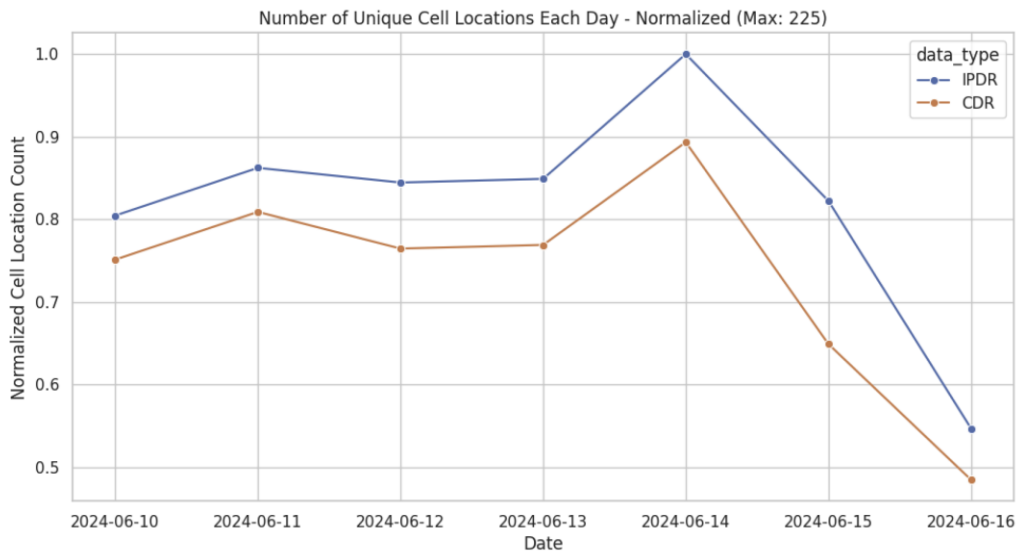
3_number_of_subscribers_daily.png

USER_CONSISTENCY_REPORT -> FAILED
Explanation:
Discrepancies or irregularities found in the daily subscriber counts, within range 1 - 98 (ratio: 1.02%).
Action plan:
Investigate the causes behind these inconsistencies, such as data processing errors, data loss, or any other factors.
Review the data collection and processing procedures to address and rectify these inconsistencies.
See attached image '3_number_of_subscribers_daily.png' for the visualization.

Number of unique cell locations

We can examine the consistency of the count of unique cell locations in the MPD dataset. Consistent counts indicate a reliable dataset, while inconsistencies may suggest data quality issues or errors in recording cell locations.

```
import matplotlib.pyplot as plt
import seaborn as sns

# Convert the subscriber counts DataFrame to Pandas DataFrame for visualization
cell_location_counts_pd = cell_location_counts_df.toPandas().sort_values('date')

cell_location_counts_pd['cell_count_norm'] = cell_location_counts_pd['cell_count']/max_count

# Plot the line chart
plt.figure(figsize=(12, 6))
sns.lineplot(data=cell_location_counts_pd, x='date', y='cell_count_norm', hue='data_type', marker="o")

# Set the chart title and labels
plt.title('Number of Unique Cell Locations Each Day - Normalized (Max: {})'.format(max_count))
plt.xlabel('Date')
plt.ylabel('Normalized Cell Location Count')

# Set the path
targeted_path_img = QA_PATH+"4_number_of_cells_daily.png"

# Save the chart as an image file at the specified location
plt.savefig(targeted_path_img)

# Show the chart
plt.show()
```



4_number_of_cells_daily.png

CELL_CONSISTENCY_REPORT -> FAILED
Explanation:
Discrepancies or irregularities found in the cell counts.
Action plan:
Investigate the reasons behind this instability.
Check for any issues related to cell tower coverage, data transmission, or data aggregation.
Collaborate with the network operations team to rectify any issues and ensure accurate and stable cell counts.

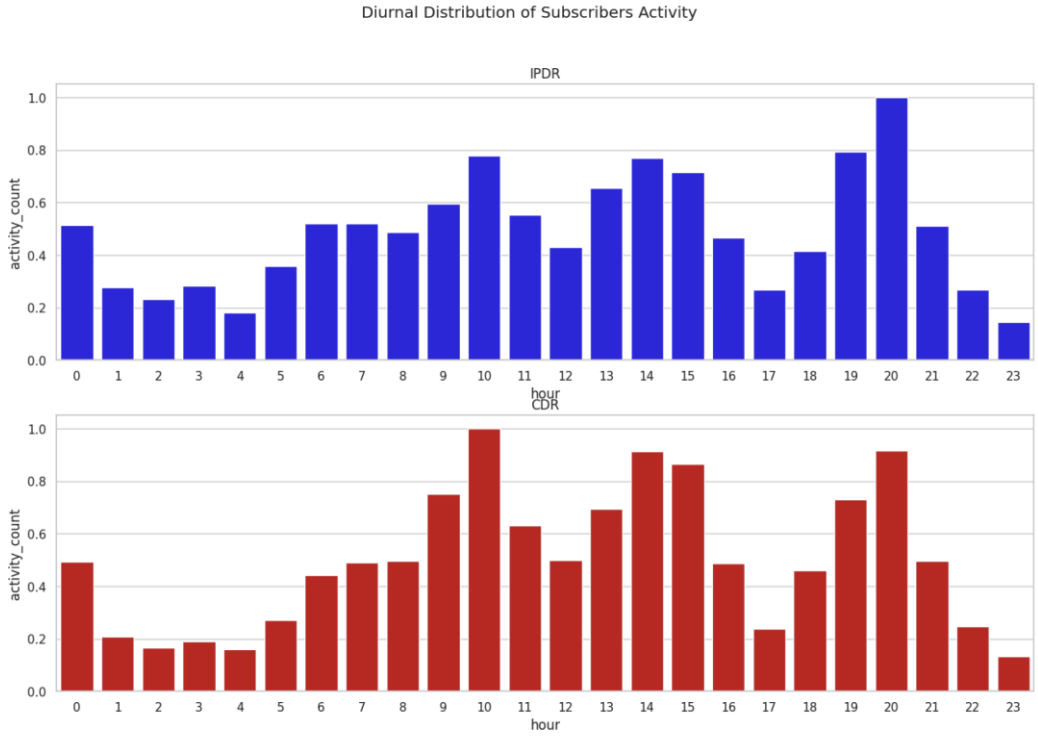
Subscribers Activity In 24 hours

The purpose of this metric is to analyze the diurnal distribution of subscribers' activity throughout the day. By examining this distribution, we can identify peak usage hours, low-usage periods, or any abnormal patterns in subscribers' activity.

Ideal distribution: In an ideal scenario, the diurnal distribution should follow an `"elephant shape"` pattern, where the activity count is highest during working hours and gradually decreases during non-working hours. This reflects typical usage patterns, with higher activity during the day and lower activity during the night.

Anomalies and patterns: Look for any anomalies or patterns that deviate from the expected distribution. These may include:

- `Flat pattern`: Unexpectedly similar counts of events in every hour indicates anomalies or abnormal data collection. It could be due to algorithmic data generation, record sampling, or other factors that drive unusual subscriber behavior.
- `Spikes or peaks`: Unexpectedly high activity counts during specific hours might indicate anomalies or abnormal patterns of usage. It could be due to events, promotions, or other factors that drive unusual subscriber behavior.
- `Inverse pattern`: Unexpectedly high activity counts during non-working hours indicating potential problem with the timezone setting in the timestamp.



diurnal_distribution.png

DIURNAL_DISTRIBUTION_REPORT -> FAILED
Explanation:
Diurnal distribution from CDR is not following the elephant shape
Diurnal distribution from IPDR is not following the elephant shape
Action plan:
Analyze the diurnal distribution patterns from CDR and IPDR data and investigate the reasons for the deviation from the expected elephant shape.
This may involve examining user behavior, network usage patterns, or potential issues with data format like incorrect timezone.
Take appropriate actions to address any discrepancies and ensure accurate diurnal distribution reporting.

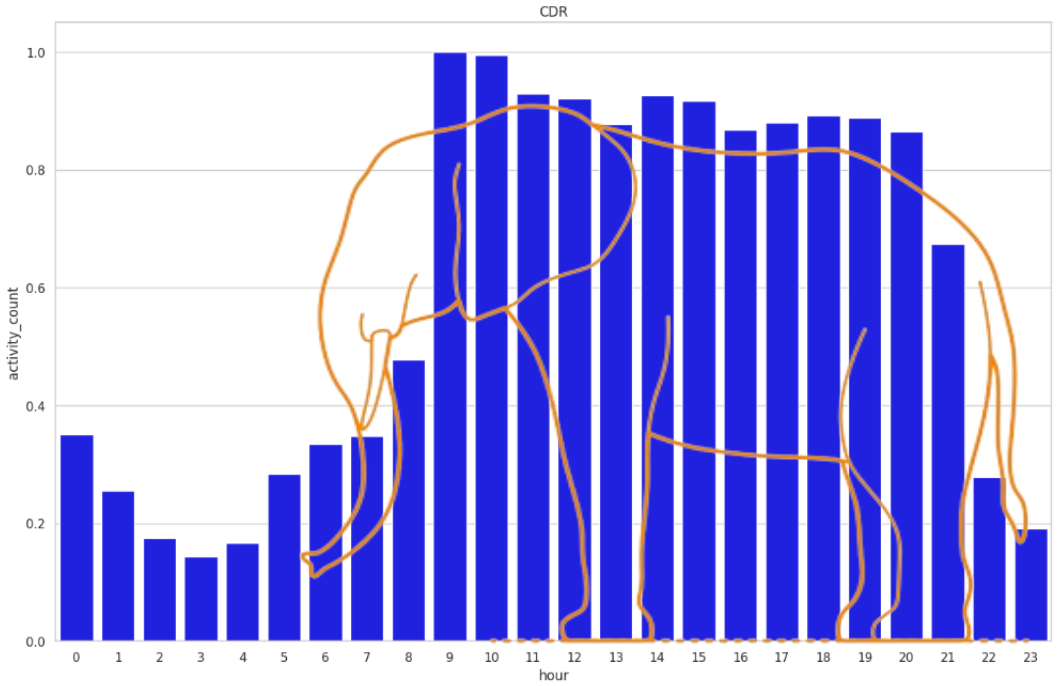
Subscribers Activity In 24 hours

The purpose of this metric is to analyze the diurnal distribution of subscribers' activity throughout the day. By examining this distribution, we can identify peak usage hours, low-usage periods, or any abnormal patterns in subscribers' activity.

Ideal distribution: In an ideal scenario, the diurnal distribution should follow an `"elephant shape"` pattern, where the activity count is highest during working hours and gradually decreases during non-working hours. This reflects typical usage patterns, with higher activity during the day and lower activity during the night.

Anomalies and patterns: Look for any anomalies or patterns that deviate from the expected distribution. These may include:

- `Flat pattern`: Unexpectedly similar counts of events in every hour indicates anomalies or abnormal data collection. It could be due to algorithmic data generation, record sampling, or other factors that drive unusual subscriber behavior.
- `Spikes or peaks`: Unexpectedly high activity counts during specific hours might indicate anomalies or abnormal patterns of usage. It could be due to events, promotions, or other factors that drive unusual subscriber behavior.
- `Inverse pattern`: Unexpectedly high activity counts during non-working hours indicating potential problem with the timezone setting in the timestamp.



diurnal_distribution.png

DIURNAL_DISTRIBUTION_REPORT -> PASSED
Explanation:
Diurnal distribution from CDR is following the elephant shape with p-value: 0.9671
Diurnal distribution from IPDR is following the elephant shape with p-value: 0.9671

4. Defining “home” location

“Home” and anchor points

- Determining a subscriber’s “home” is a crucial step for MPD uses cases, *e.g.* commuting, information society etc.
- “Home” is used to map MPD with reference data, e.g. LAU and population
- A home location in an LAU is assumed to be included in the population estimate for that LAU.
- Other methods include to define more ‘anchor’ points where people regularly stay, e.g. “work”

Many ways to define 'home'

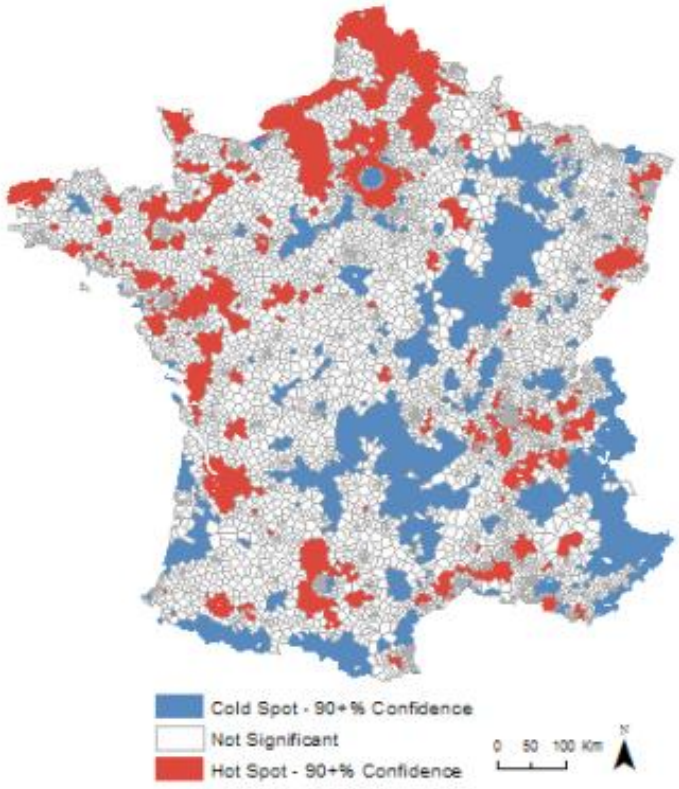
- The `amount of activity` – "home" is defined as the cell location from where most calls and texts were recorded
- The `number of active days` – "home" is the cell location from where calls and texts were recorded on the highest number of distinct days
- `Time constraints` – "home" is the cell location from where most calls and texts were recorded between 7 p.m. and 9 a.m.
- `Spatial aggregation` – "home" is the cell location from where most calls and texts were recorded within a spatial perimeter, e.g. 1km, around a cell and aggregating all activities within that perimeter
- `Combination` of time constraints and spatial aggregation.
- More `sophisticated models` also developed.

Validating ‘home’ algorithms and how criteria used influence the results

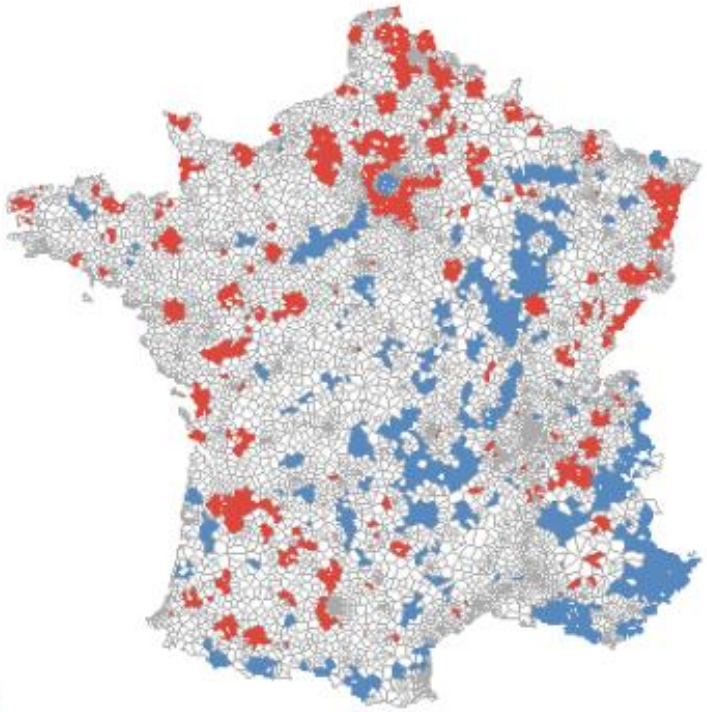
- In Estonia, a “home” anchor point model was up to 99% accurate at the county level and over 90% for higher-level LAU.
- A study of five home algorithms in France showed that the criteria used influenced the detection of home locations for up to about 40% of subscribers.

Avoid summer / vacation season

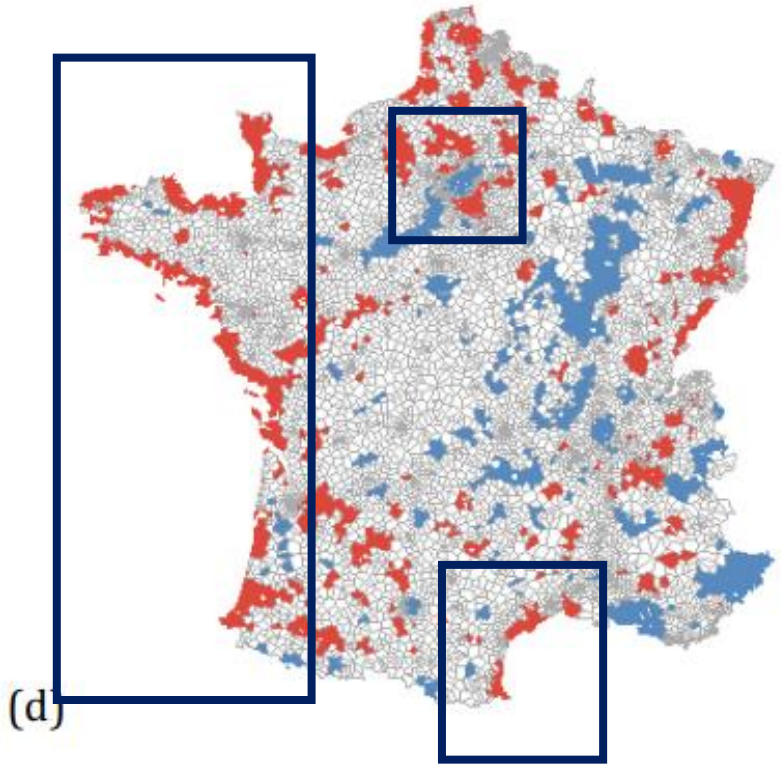
Validation



June



August



Maarten Vanhoof et al., Assessing the quality of home detection from mobile phone data for official statistics

‘Home’ algorithm used in the ITU codes

- Based on Brazil study: a time-constrained “anchoring” model (used to calculate *SDG indicator 17.8.1 Proportion of individuals using the Internet)*
- The primary objective is to infer the most probable cell location that can be considered the subscriber's home location.
- “Home cell location” identified by:
  - the number events at each cell location from Mondays to Thursdays
  - prioritization within three specific times (Night, Morning, Evening)

User summary - aggregation by subscribers

- The data is summarized into two sets of summary data: `User Summary` and `Cell Summary`.
- `User Summary` calculates user activity statistics, *e.g.* the number of events/records, unique cell locations, unique days, principal technology used, and highest technology used
- It is needed for indicator calculation to determine whether the user has used the Internet, how often, and by which technology.

msisdn	internet_user	IPDR_events	CDR_events	IPDR_unique_cell	CDR_unique_cell
31	TRUE	635	202	48	15
85	TRUE	2,332	237	76	14
65	TRUE	2,681	73	82	14
53	TRUE	2,910	174	50	10
78	TRUE	1,732	266	61	16

Cell summary - identifying the “home cell”

1. Each event (CDR, IPRD) is classified according to four anchor time categories “Night", "Morning","Night“ and "Office Hours".

Default settings:

- Anchor #1 : (00 - 05) -> “Night"
- Anchor #2 : (05 - 08) -> "Morning"
- Anchor #3 : (21 - 00) -> "Night"
- Outtime: (08 - 21) -> "Office Hours"

```
filtered_cell = cell_stats.filter((cell_stats.msisdn == 'subscribers_00007') & (cell_stats.date == '2024-06-10'))
filtered_cell.show()
```

datetime	cell_id	latitude	longitude	data_type	service	date	msisdn	hou	anchor_type
2024-06-10 21:20:00	2119	43.303	-3.036	CDR	3G	2024-06-10	subscribers_00007	2	ANCHOR_3
2024-06-10 19:04:00	2119	43.303	-3.036	IPDR	3G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 19:57:00	2119	43.303	-3.036	CDR	3G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 05:30:00	2119	43.303	-3.036	CDR	3G	2024-06-10	subscribers_00007		ANCHOR_2
2024-06-10 10:28:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 09:46:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007		OUTTIME
2024-06-10 13:36:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 12:26:00	528	43.262	-2.942	IPDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 15:56:00	528	43.262	-2.942	IPDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 11:19:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 14:46:00	528	43.262	-2.942	IPDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 10:42:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 08:51:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007		OUTTIME
2024-06-10 08:20:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007		OUTTIME
2024-06-10 14:13:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 15:20:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 07:15:00	528	43.262	-2.942	IPDR	2G	2024-06-10	subscribers_00007		ANCHOR_2
2024-06-10 13:13:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME

Cell summary

2. Aggregate number of events per subscriber, cell and anchor per day

msisdn	date	anchor_type	cell_id	cnt
subscribers_00007	2024-06-10	OUTTIME	528	13
subscribers_00007	2024-06-10	ANCHOR_2	2119	1
subscribers_00007	2024-06-10	OUTTIME	2119	2
subscribers_00007	2024-06-10	ANCHOR_2	528	1
subscribers_00007	2024-06-10	ANCHOR_3	2119	1

3. Assign the most used cell for each anchor and day

msisdn	date	is_weekday	anchor_type	cell_id	cnt
subscribers_00007	2024-06-10	true	ANCHOR_2	2119	1
subscribers_00007	2024-06-10	true	ANCHOR_2	528	1
subscribers_00007	2024-06-10	true	ANCHOR_3	2119	1
subscribers_00007	2024-06-10	true	OUTTIME	528	13

Cell summary

4. For each subscriber – aggregate the number of days for which the cell is the defined ‘anchor time’

msisdn	is_weekday	anchor_type	cell_id	day_cnt
subscribers_00007	true	ANCHOR_1	2119	3
subscribers_00007	true	ANCHOR_2	2119	3
subscribers_00007	true	ANCHOR_2	528	2
subscribers_00007	true	ANCHOR_3	2119	4
subscribers_00007	true	OUTTIME	528	4

5. Infer “home” cell according to multi-step logic

“Direct inference” → “Tiebreaker” → “Indirect inference”

Determining 'home cell' #1/3

1. "Direct inference"

- If the subscriber has one dominant cell during most days during Anchor time 1 (0-5 am), the code assigns the cell to be the subscriber's "home cell".
- If the subscriber has two or more cells with the same day count, no "home cell " is assigned.

Example:

- Out of 30 days, CellID = 123 is the most frequently used cell for Anchor time 1.

CellID = 123 is assigned to be the subscriber's "home cell"

Determining ‘home cell’ #2/3

2. “Tiebreaker”

- If one of the cells identified during Anchor time 1 is the dominant cell during Anchor time 2 (5-8am), the cell is assigned to be the subscriber’s “home cell”
- If none, the Anchor time 3 (9pm-0am) is checked.
- If multiple other cells are identified to be dominant during Anchor time 2 and Anchor time 3, no “home cell” is assigned.

Example:

- Subscriber “A” has the same number of day count in two cell IDs for Anchor time 1: CellID = 101 and CellID = 936.
- For Anchor time 2, CellID = 936 has more day count than CellID = 101

CellID = 936 is assigned to be subscriber “A”’s “home cell”

Determining ‘home cell’ #3/3

3. “Indirect inference”

- If the subscriber’s “home cell” still can’t be identified by direct or tiebreaker inference, the code randomly picks one location from all the highest frequency candidate locations by prioritizing **Anchor time 1 > Anchor time 2 > Anchor time 3.**

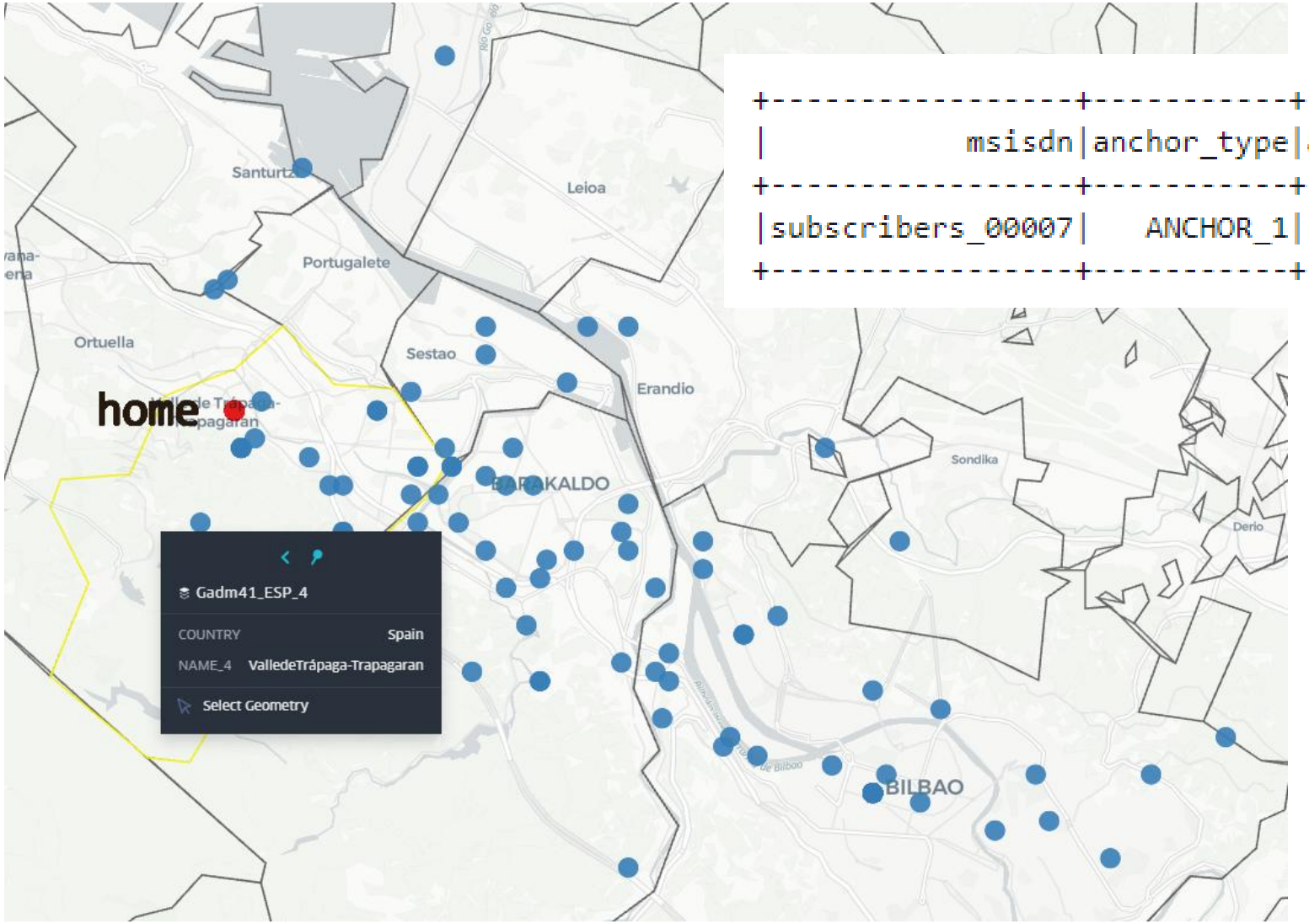
Example:

The anchoring result for subscriber “A” is described below:

- Anchor time 1: [{cell_id: “202”, days: 20}, {cell_id: “303”, days: 20}]
- Anchor time 2: [{cell_id: “505”, days: 18}, {cell_id: “404”, days: 18}]
- Anchor time 3: No Data

CellID = 202 from Anchor time 1 is chosen as subscriber “A”’s “home cell”

Example



msisdn	anchor_type	anchor_cell_id	anchor_label
subscribers_00007	ANCHOR_1	2119	DIRECT

5. Indicator calculation and next steps

What is indicator 17.8.1 - recall

- Goal 17: Strengthen the means of implementation and revitalise the global partnership for sustainable development

- Target 17.8: Enhance the use of enabling technology

- **Indicator 17.8.1: Proportion of individuals using the Internet from any location in the past 3 months**

What is indicator 17.8.1 - recall

- Proportion of people using mobile phone data (Internet)
- Breakdown by technology: 2G, 3G, 4G/LTE, 5G
- Breakdown by geography: local administrative units (LAU)
- Where subscribers live (home anchor)
- What is the dominant/highest technology used

Steps for calculation

1. **Define the Scope of Aggregation:** Determine the geographic level (global, regional, national, or local) and specify the partnerships or initiatives to be considered for measurement.
2. **Identify Data Sources:** Join information from anchoring process with another data source like Customer relationship management (CRM) data that contains information about subscribers demography, and other data relevant data sources.
3. **Calculate the Indicator:** Use the formula to calculate the indicator:

   Indicator 17.8.1 = (Number of subscribers using the Internet/ Total number of individuals in the target population) x 100%

4. **Data Validation:** Validate the accuracy and reliability of the calculated indicator from mobile data by comparing it with other sources such as household survey to ensure its quality and coherence.

Calculation of indicator 17.8.1 using MPD

- **Sum of subscribers** with voice only events, data (Internet) only events, and both (voice + Internet) events

- **Total no. of people using Internet**. That is subscribers with:
  - Data only events
  - Both data and voice events

- **Enrich by adding technology breakdown.** That is by selecting most frequently used technology (2G/3G/4G) by a subscriber for Internet access.

Method for calculating 17.8.1

To calculate this proportion, aggregated data from the MNO is used.

The **proportion of individuals using the Internet** will be calculated with the following equation:

$$proportion\ of\ people\ using\ internet\ (lau2) = \frac{data\ users\ home\ count}{home\ count}$$

EXAMPLE:

$$\text{proportion of people using 3G} = \frac{\text{3G\_home\_count}}{\text{total\_home\_count}}$$

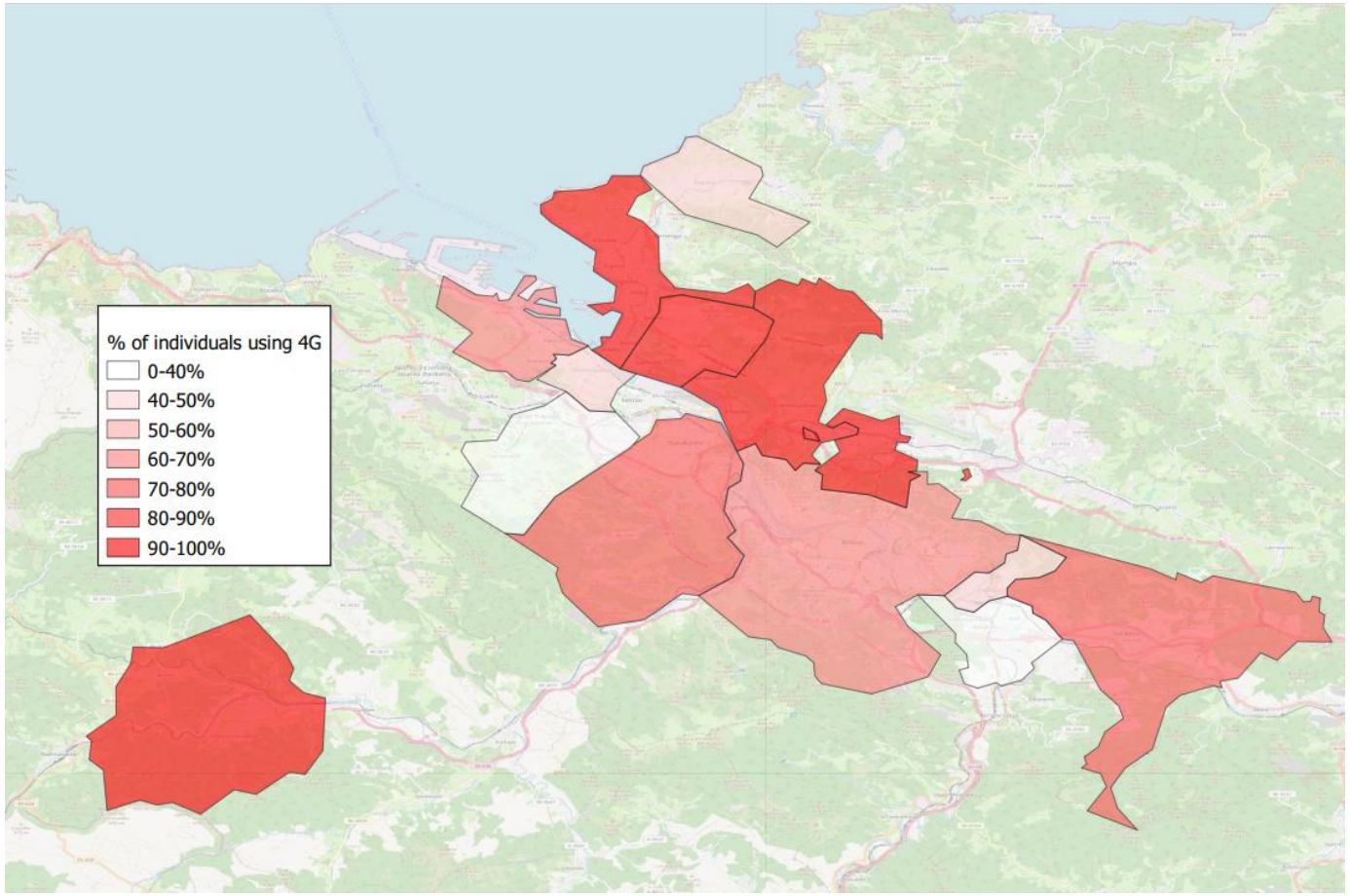
Example tables

In [11]:
```
# Define the column name to be used in the aggregation
tech_col = 'IPDR_highest_service'

# Group the dataframe 'df' by the 'municipality' column and perform aggregations
ind_adm = df.groupBy(['municipality']).agg(
        count(col('msisdn')).alias('total_home_count'),  # Count the number of 'msisdn' values and alias the column as 'total_
        count(when(col('internet_user')==False, col('msisdn'))).alias('total_no_internet'),  # Count the number of 'msisdn' va
        count(when(col('internet_user')==True, col('msisdn'))).alias('total_internet_ALL'),  # Count the number of 'msisdn' va
        count(when((col('internet_user')==True) & (col(tech_col)=='2G'), col('msisdn'))).alias('total_internet_2G'),  # Count
        count(when((col('internet_user')==True) & (col(tech_col)=='3G'), col('msisdn'))).alias('total_internet_3G'),  # Count
        count(when((col('internet_user')==True) & (col(tech_col)=='4G'), col('msisdn'))).alias('total_internet_4G')  # Count t
    ).orderBy(['municipality']).toPandas()
ind_adm
```

Out[11]:

	municipality	total_home_count	total_no_internet	total_internet_ALL	total_internet_2G	total_internet_3G	total_internet_4
0	Barakaldo	8	0	8	0	2	
1	Basauri	3	0	3	0	2	
2	Bilbao	56	0	56	5	12	3
3	Erandio	2	0	2	0	0	
4	Etxebarri,AnteiglesiadeSanEs	2	0	2	0	1	
5	Galdakao	4	0	4	0	1	
6	Getxo	1	0	1	0	0	
7	Leioa	1	0	1	0	0	
8	Portugalete	6	0	6	0	3	
9	Santurtzi	3	0	3	0	1	
10	Sondika	1	0	1	0	0	
11	Sopelana	2	0	2	0	1	
12	ValledeTrápaga-Trapagaran	3	0	3	0	2	
13	Zalla	1	0	1	0	0	

Example: Bilbao area, using synthetic MPD (% of individuals using 4G)



Calculating Indicator 17.8.1

Utilizing Mobile Positioning Data (MPD) to Measure SDGs Indicator 17.8.1: Proportion of Individuals Using the Internet

Summary

`SDG Indicator 17.8.1` aims to assess the proportion of individuals using the Internet, specifically those who have accessed it within the last three months. The Internet, being a global computer network that facilitates various communication services like the World Wide Web, email, news, and entertainment, can be accessed through different devices such as computers, mobile phones, tablets, PDAs, gaming consoles, and digital TVs. Connectivity may be established through fixed or mobile networks.

Leveraging mobile phone data offers a valuable avenue for estimating and enhancing the measurement of this indicator. The utilization of mobile phone data can significantly contribute to the accurate and detailed assessment of SDG Indicator 17.8.1 by providing real-time and granular information on Internet usage. However, it is essential to approach data privacy, data quality, and the need for complementary data sources with careful consideration to ensure the reliability and effectiveness of the measurement process.

When feasible, the indicator can be further analyzed by various breakdowns and disaggregation criteria, such as regional distinctions (urban and rural areas), gender, age groups, etc. The International Telecommunication Union (ITU) collects data from countries encompassing these breakdowns to enhance the comprehensiveness of the indicator's assessment.

Calculation of Indicator

add highlight for hashing consistency from MNO

To obtain geographical coordinates and facilitate data aggregation, it is essential to combine Mobile Positioning Data (MPD) with cell data obtained from Mobile Network Operators (MNOs) or other publicly available sources. The cell data includes approximate geographical coordinates and the corresponding technology used. Each cell's location, categorized as the subscriber's home location, is then aggregated at the desired geographical level, such as LAU 2, and separated based on the technology used for internet connectivity.

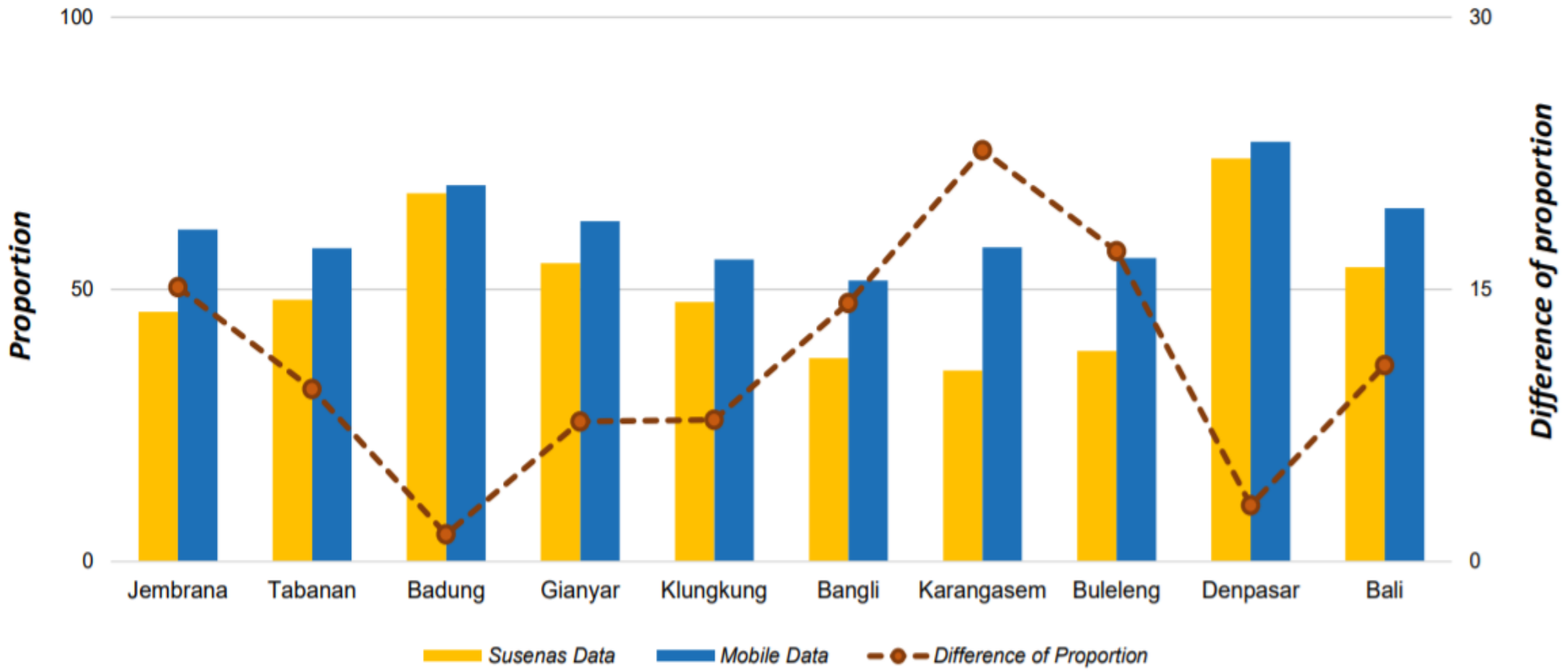
The following assumptions are made regarding subscriber internet usage:

- If a subscriber has no mobile internet traffic (IPDR), it is assumed that they do not have internet access.
- If a subscriber only has mobile internet traffic in 2G cells (without any events in 3G or 4G), it is presumed that they solely

Example:

	INTERNET ACCESS (%)		INTERNET TECHNOLOGY (%)		
MUNICIPALITY	**NO**	**YES**	**2G**	**3G**	**4G**
Belford Roxo	4.78	95.22	2.07	40.47	57.46
Cachoeira de Macacu	9.03	90.97	6.74	54.01	39.25
Duque de Caxias	3.83	96.17	0.98	37.89	61.12
Engenheiro Paulo de Frontin	6.68	93.32	30.87	59.58	9.55
Guapimirim	3.30	96.70	1.34	64.49	34.17
Itaboraí	7.47	92.53	5.08	38.74	56.18

Comparison of MPD & survey data from Indonesia



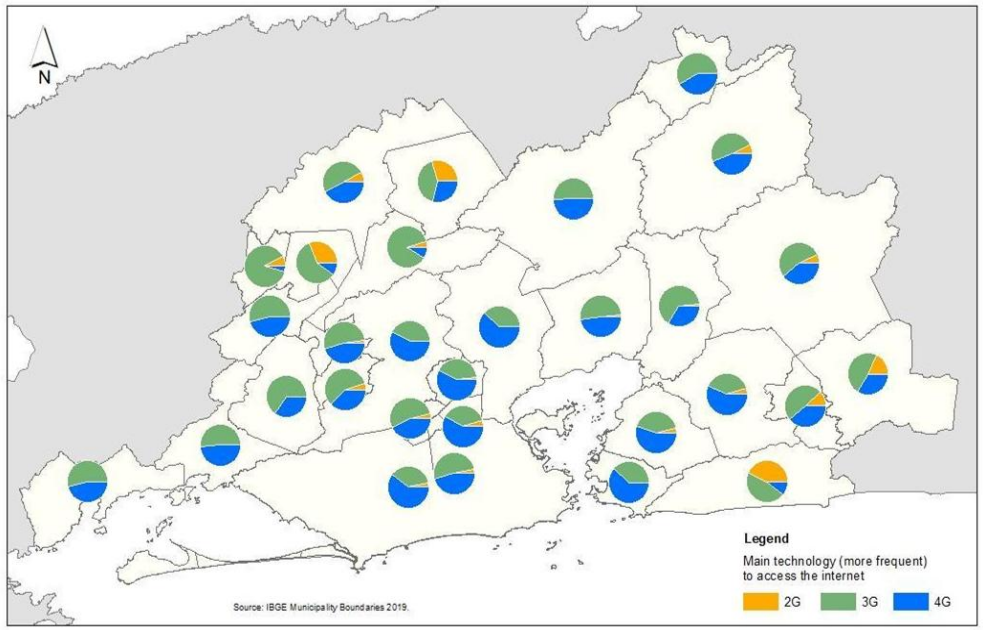
Comparison of MPD & survey data from Rio de Janeiro, Brazil

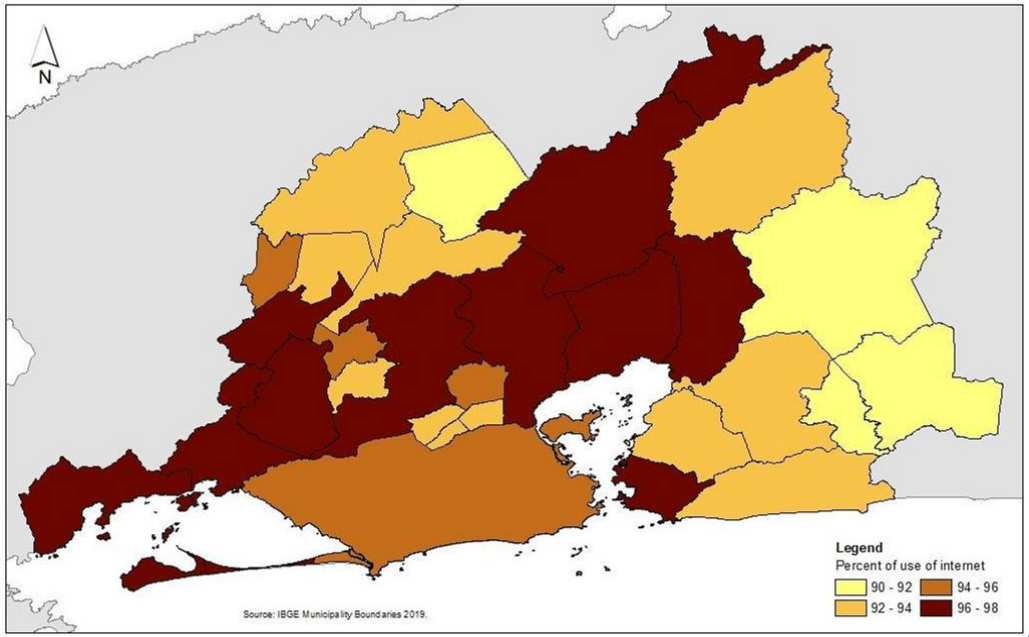
Geographic disaggregation	Internet use (%)		
	Mobile phone data	Survey data	Difference
Area of study	93.91	93.89	0.02 p.p.
Rio de Janeiro Metropolitan Area	95.04	94.01	1.04 p.p.
City of Rio de Janeiro	94.87	95.57	-0.70 p.p.

Source: Mobile data and Q4 2018 PNAD Contínua Survey/IBGE.

Example from Brazil

The Internet usage is high in all municipalities, but it is lower in those areas where people use mainly older generation mobile network technologies.





Thank you very much!

https://www.itu.int/en/ITU-D/Statistics/Pages/bigdata