Before working with MPD

Methodological decisions to make

Siim Esko

Positium

Estonia

@positium
www.positium.com

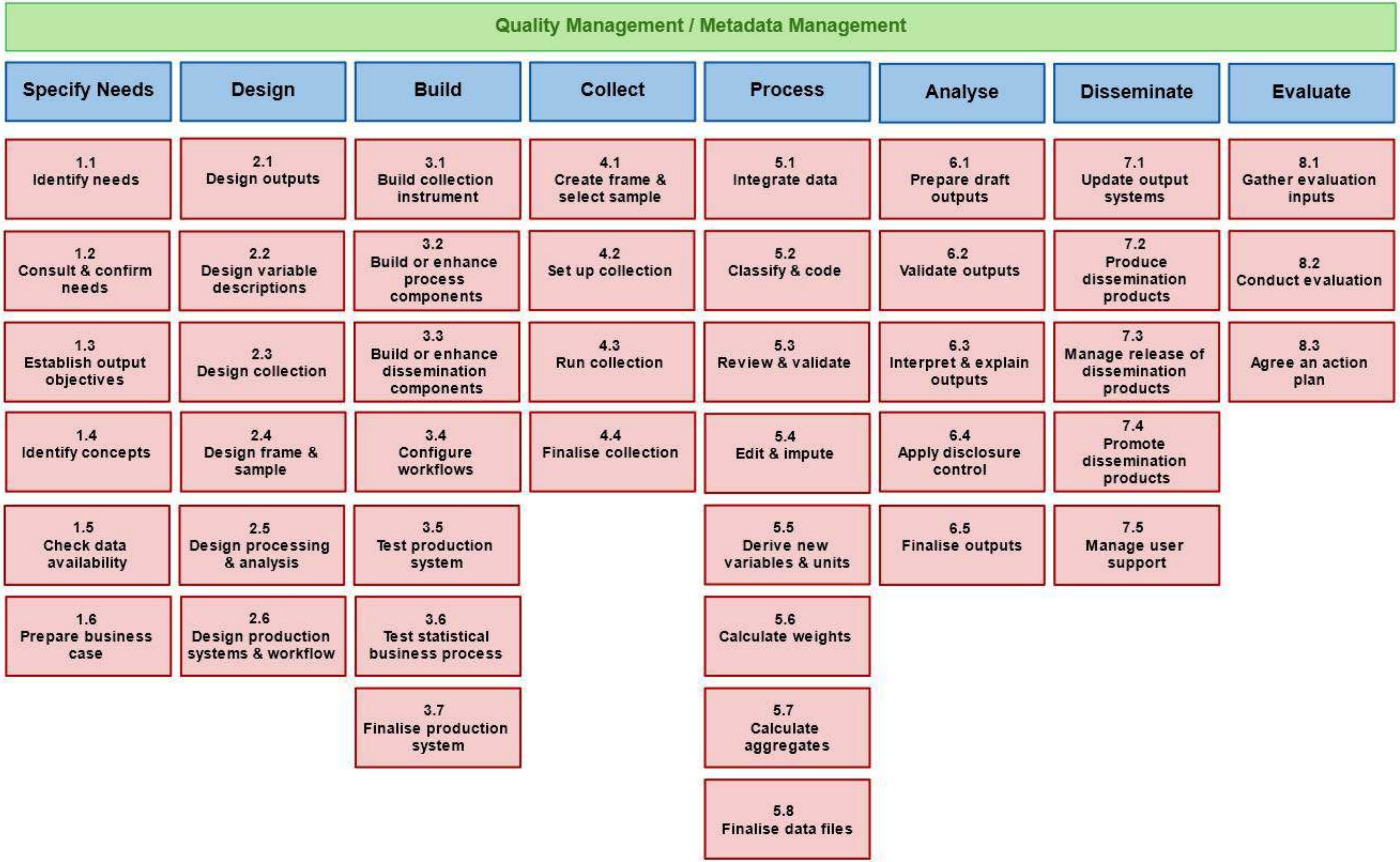
Activating question

Assume you just got access to mobile phone data. What are the first questions you have in your mind before you put a team to work on the data?

- Big Data in GSBPM -

What can NSO do themselves

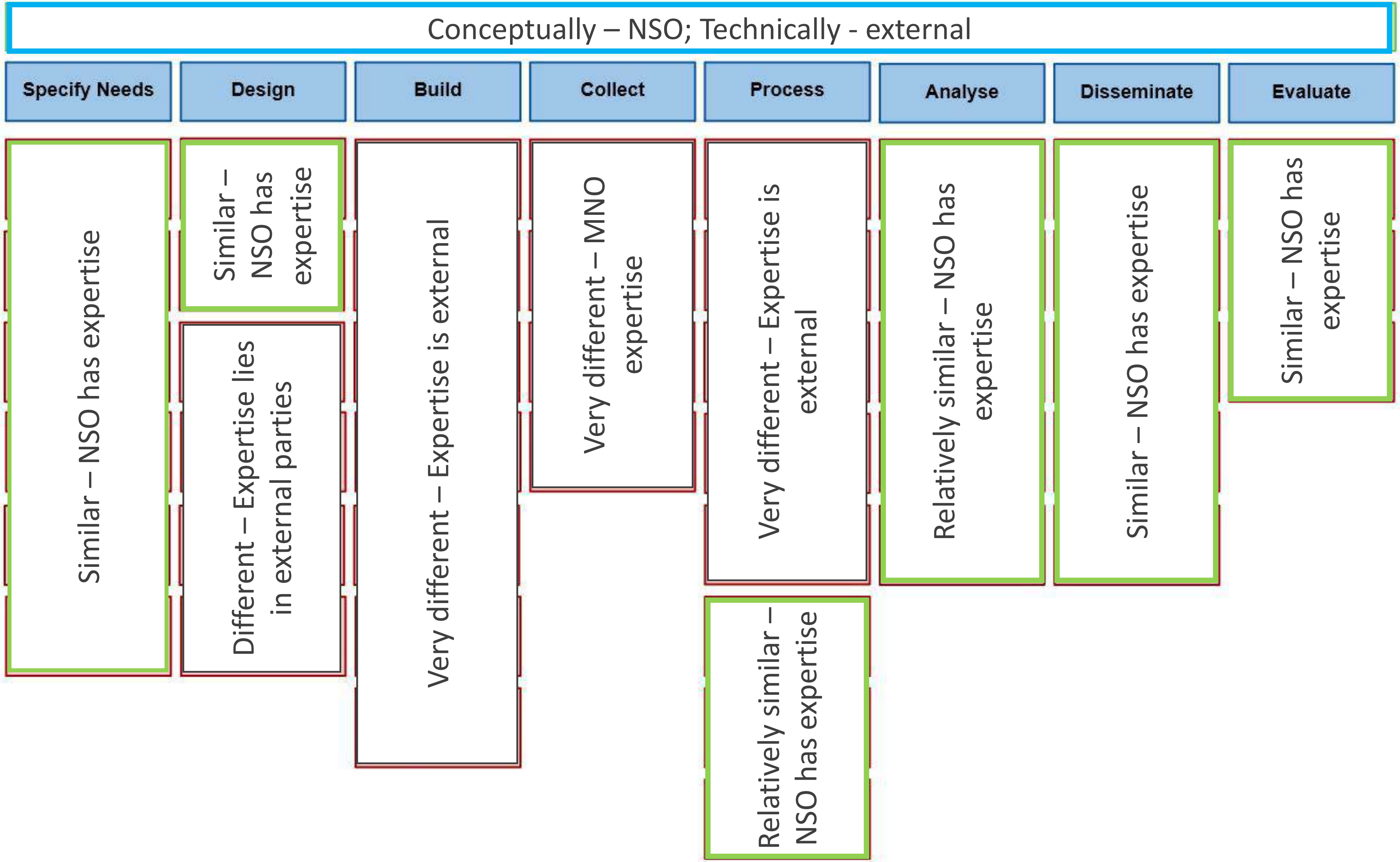
GSBPM 5.0



Quality Management / Metadata Management

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			

GSBPM new areas for NSO in terms of big data

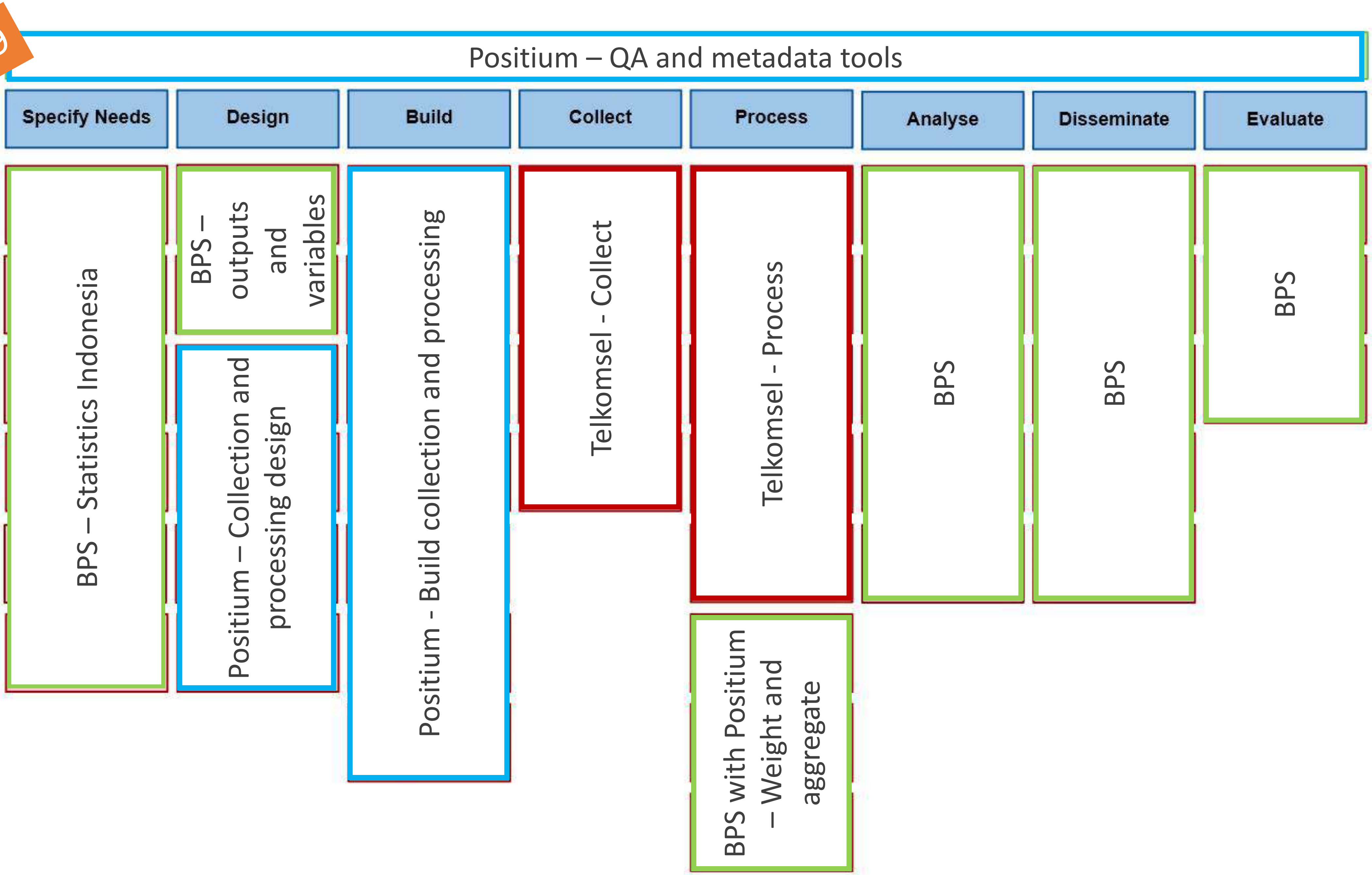


Conceptually – NSO; Technically - external

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
Similar – NSO has expertise	Similar – NSO has expertise	Very different – Expertise is external	Very different – MNO expertise	Very different – Expertise is external	Relatively similar – NSO has expertise	Similar – NSO has expertise	Similar – NSO has expertise
	Different – Expertise lies in external parties			Relatively similar – NSO has expertise			

GSBPM for cross–border tourism processing

2019



Positium – QA and metadata tools

Specify Needs

BPS – Statistics Indonesia

Design

BPS – outputs and variables

Positium – Collection and processing design

Build

Positium - Build collection and processing

Collect

Telkomsel - Collect

Process

Telkomsel - Process

BPS with Positium – Weight and aggregate

Analyse

BPS

Disseminate

BPS

Evaluate

BPS

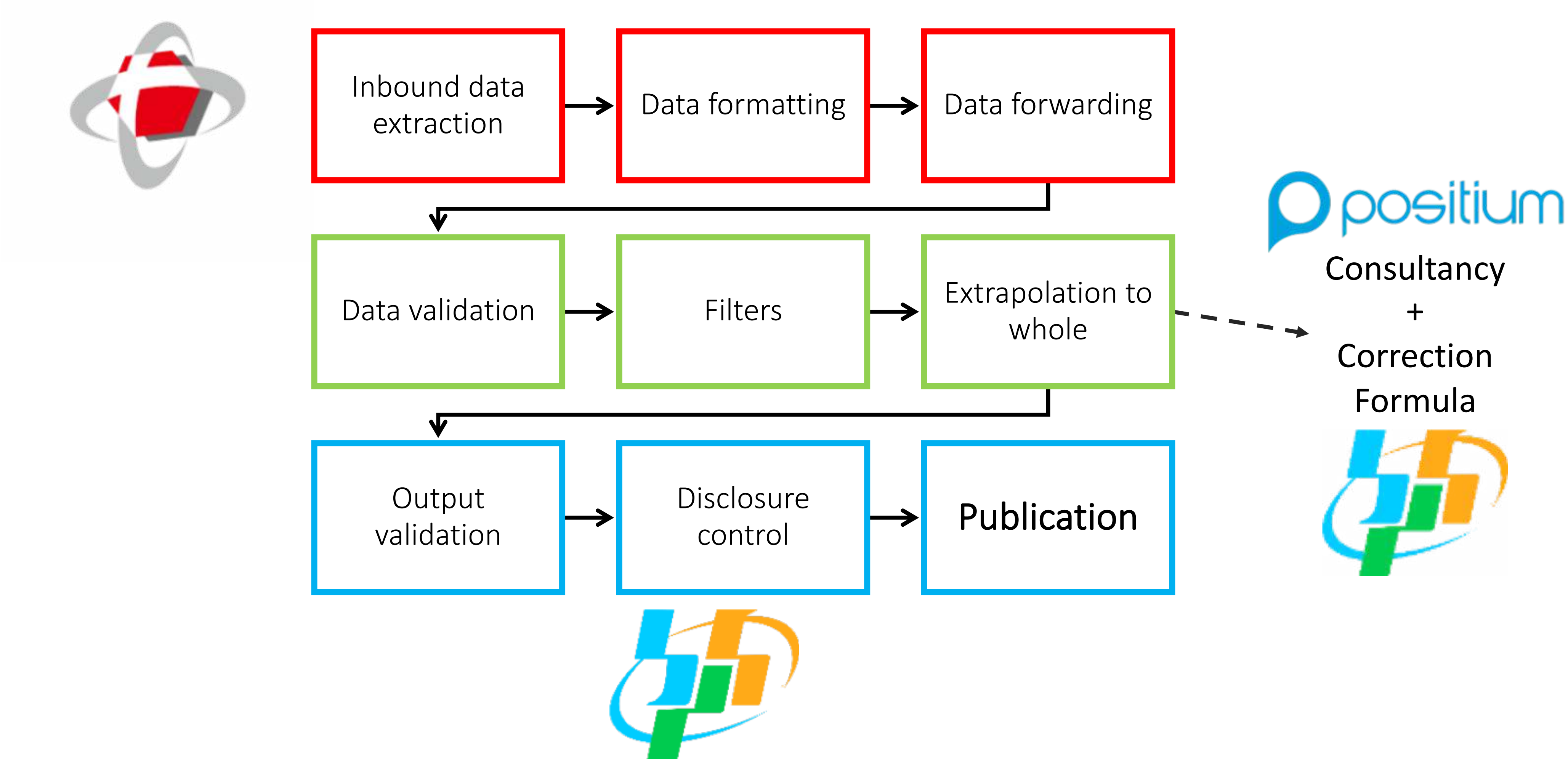
- Simple or complex are options -

Start simple or invest in advance

Basic options for methodology

Simple	**Complex**
Aggregating the data “as is” or with simple filters	Work on raw data
Remove coverage issues with calibration	Remove coverage issues with algorithms

Indonesia – Cross–border tourism – 2018



Inbound data extraction

Data formatting

Data forwarding

Data validation

Filters

Extrapolation to whole

Output validation

Disclosure control

Publication

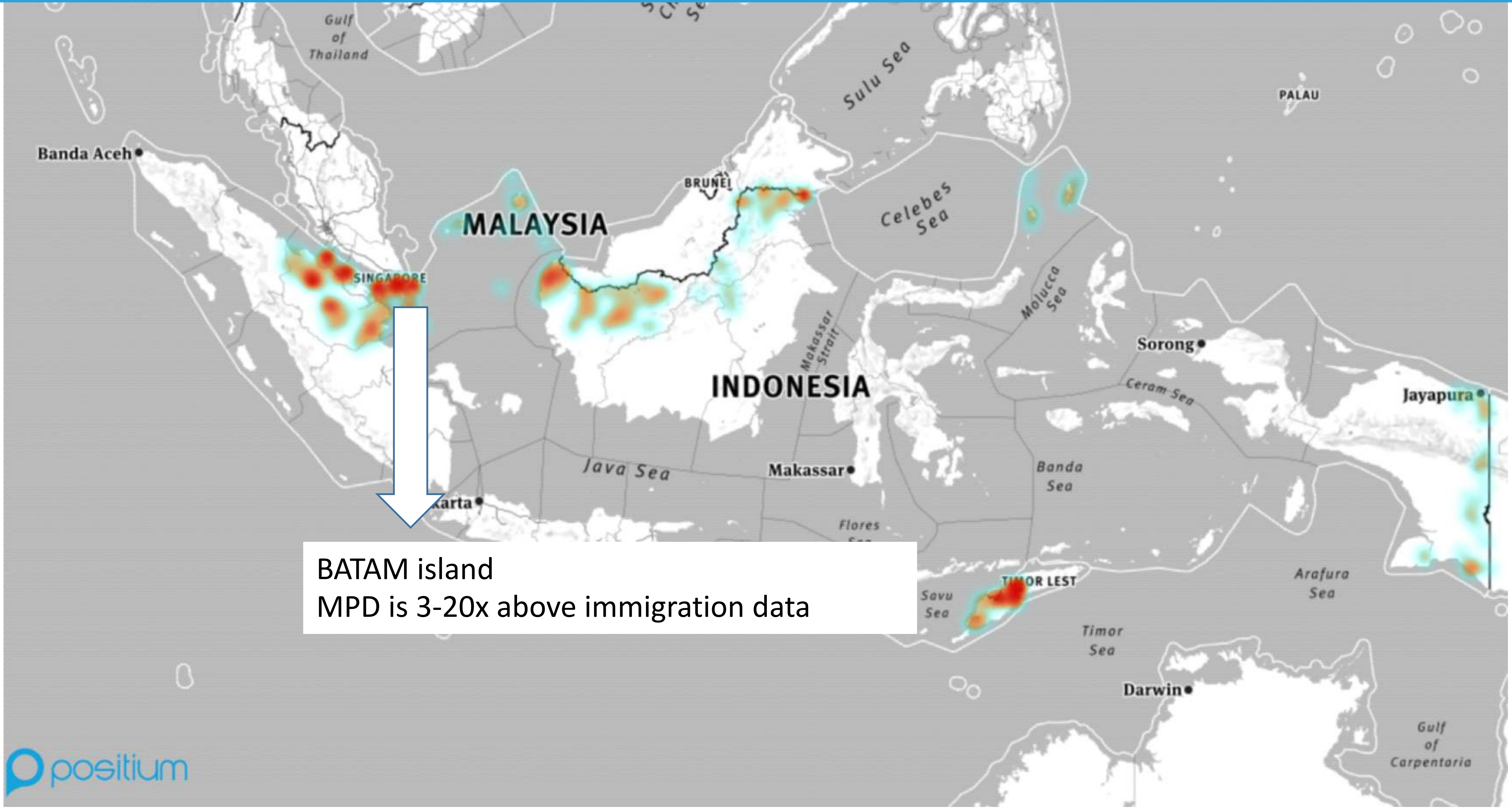
positium

Consultancy

+

Correction Formula

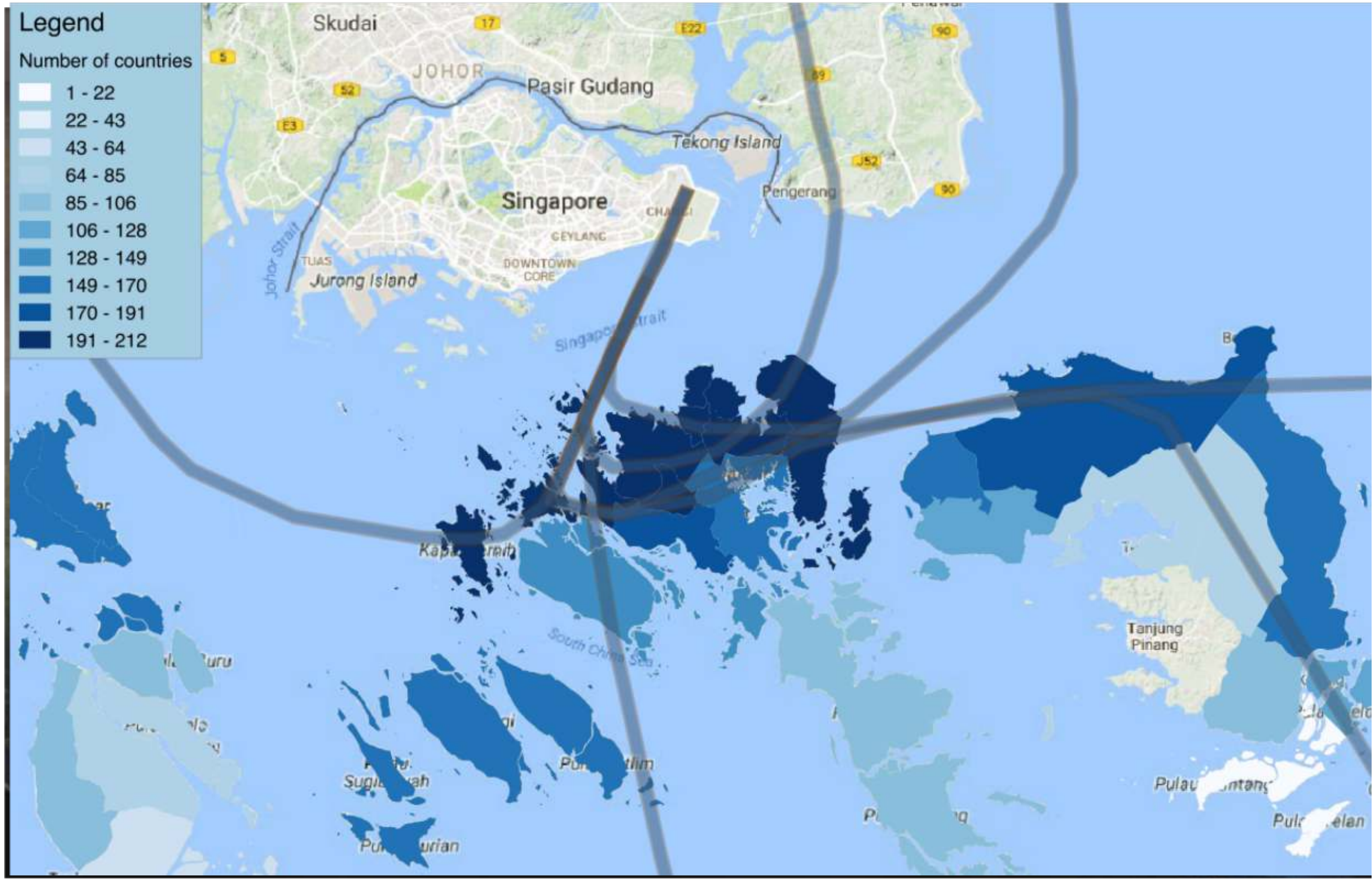
Case of Indonesia – Cross–border tourism



BATAM island
MPD is 3-20x above immigration data

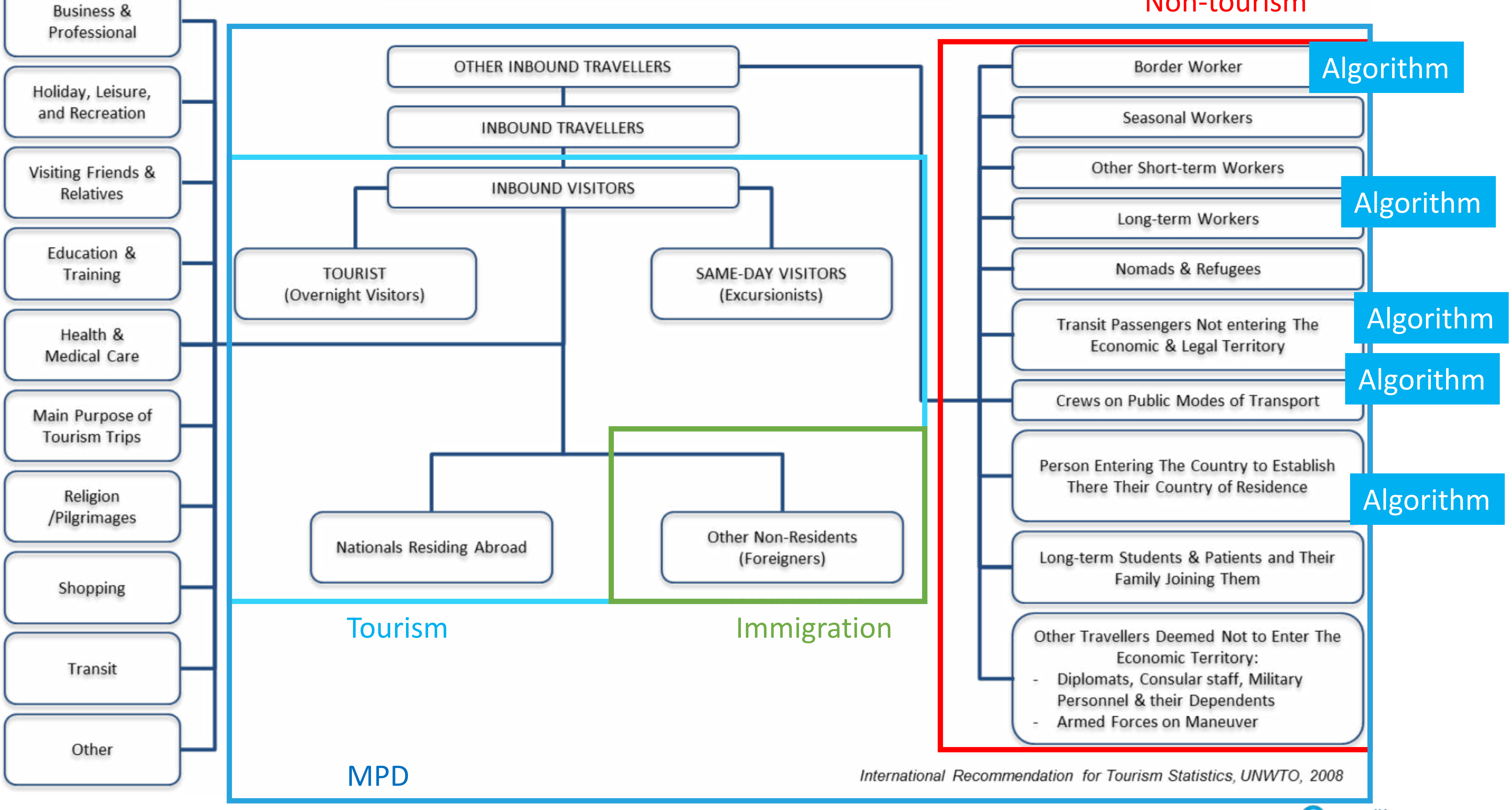
Transit: Flights

- Fast movers = those who cover the distance between two BTS that is only possible on a plane
- % of fast movers is very high in Batam and Bintan
- Do not enter the economic territory of Indonesia
- The bias exists all over Indonesia
- Can be countered by excluding fast movers



Flight path from Changi airport correlates with MPD anomalies

CLASSIFICATION OF INBOUND TRAVELLERS



Non-tourism

Business & Professional

Holiday, Leisure, and Recreation

Visiting Friends & Relatives

Education & Training

Health & Medical Care

Main Purpose of Tourism Trips

Religion /Pilgrimages

Shopping

Transit

Other

OTHER INBOUND TRAVELLERS

INBOUND TRAVELLERS

INBOUND VISITORS

TOURIST (Overnight Visitors)

SAME-DAY VISITORS (Excursionists)

Nationals Residing Abroad

Other Non-Residents (Foreigners)

Tourism

Immigration

MPD

Border Worker

Algorithm

Seasonal Workers

Other Short-term Workers

Algorithm

Long-term Workers

Nomads & Refugees

Transit Passengers Not entering The Economic & Legal Territory

Algorithm

Crews on Public Modes of Transport

Algorithm

Person Entering The Country to Establish There Their Country of Residence

Algorithm

Long-term Students & Patients and Their Family Joining Them

Other Travellers Deemed Not to Enter The Economic Territory:

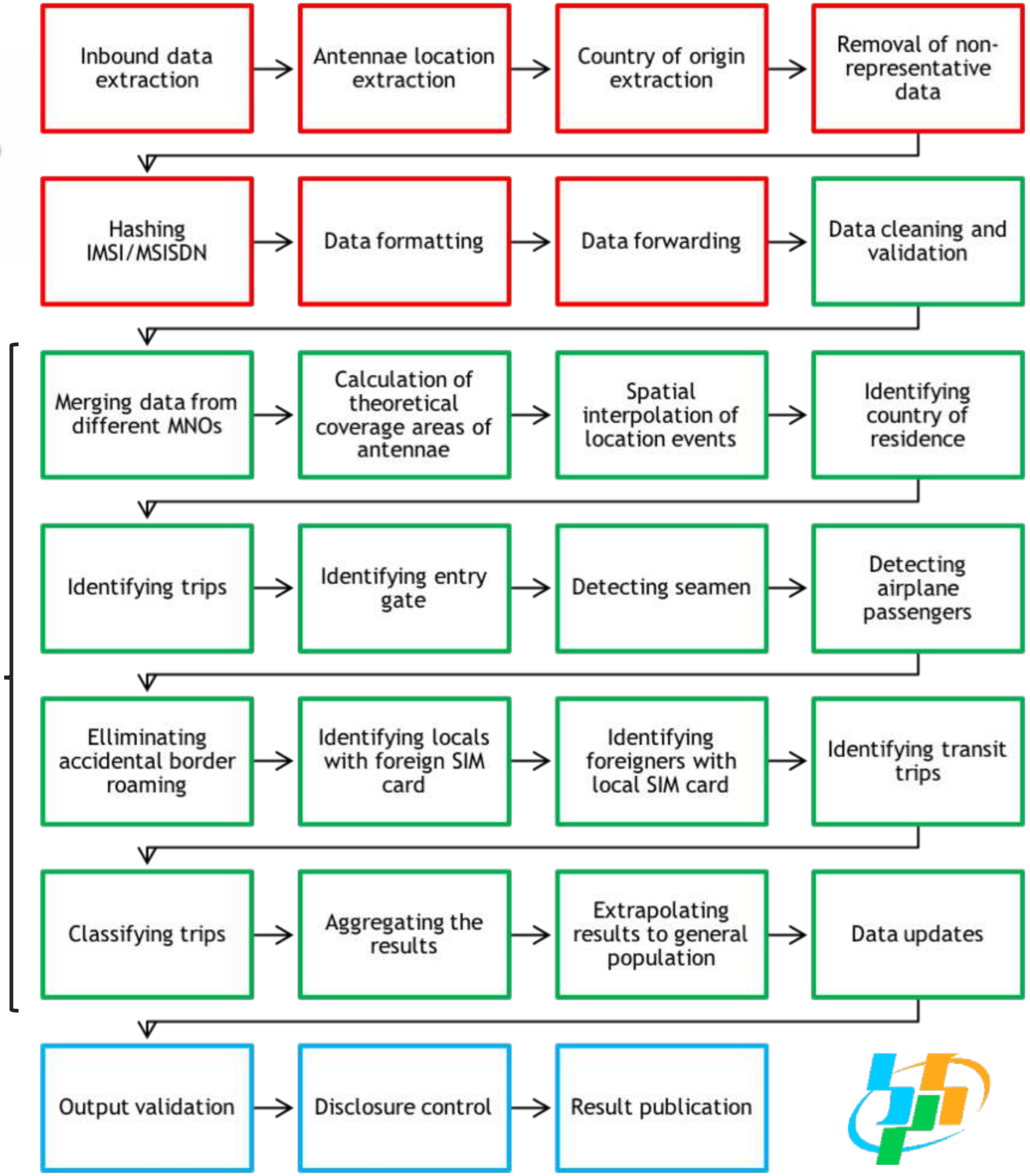
- Diplomats, Consular staff, Military Personnel & their Dependents
- Armed Forces on Maneuver

International Recommendation for Tourism Statistics, UNWTO, 2008

positium

Processing 2019

positium at



Inbound data extraction → Antennae location extraction → Country of origin extraction → Removal of non-representative data

Hashing IMSI/MSISDN → Data formatting → Data forwarding → Data cleaning and validation

Merging data from different MNOs → Calculation of theoretical coverage areas of antennae → Spatial interpolation of location events → Identifying country of residence

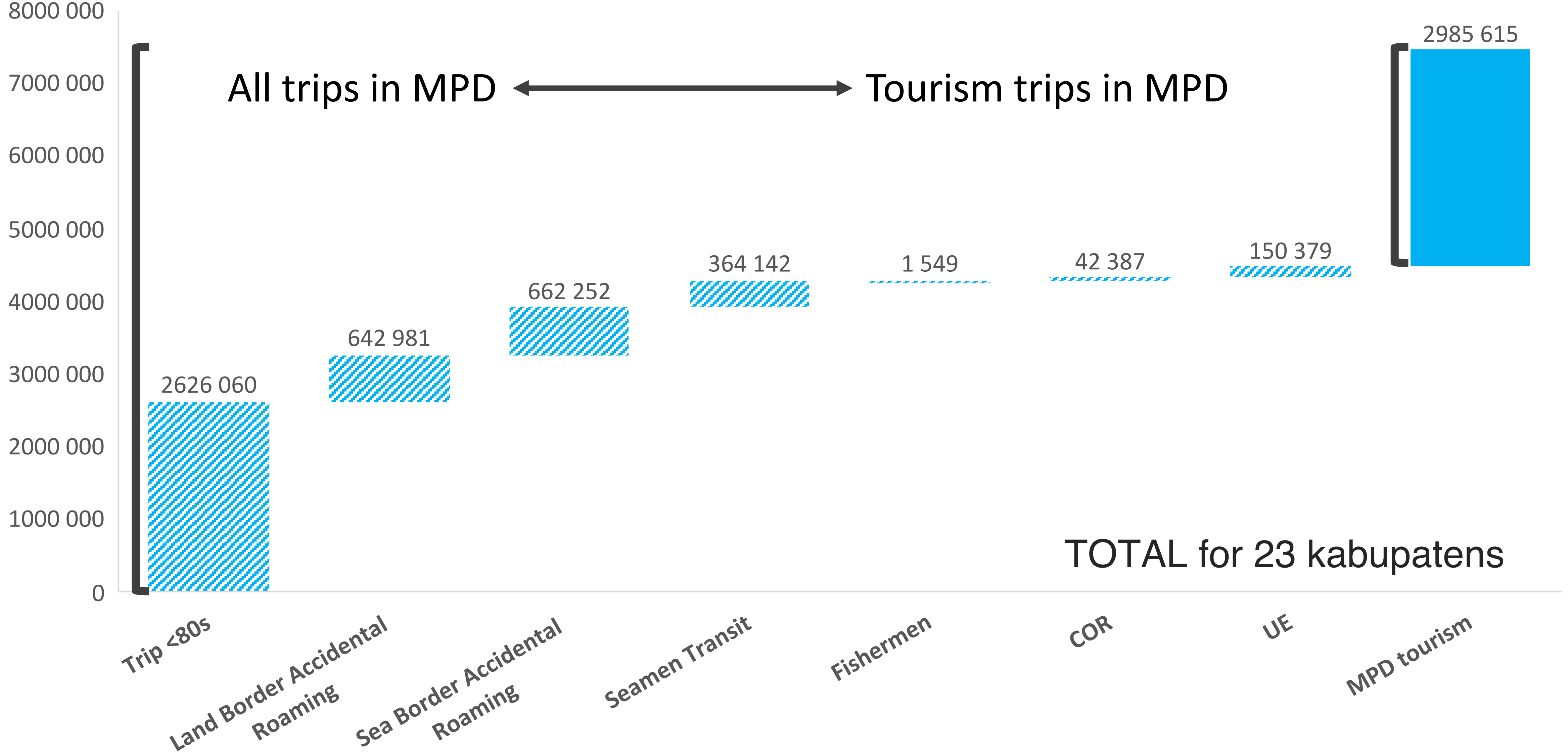
Identifying trips → Identifying entry gate → Detecting seamen → Detecting airplane passengers

Elliminating accidental border roaming → Identifying locals with foreign SIM card → Identifying foreigners with local SIM card → Identifying transit trips

Classifying trips → Aggregating the results → Extrapolating results to general population → Data updates

Output validation → Disclosure control → Result publication

Cascading of MPD data across error classes, one year



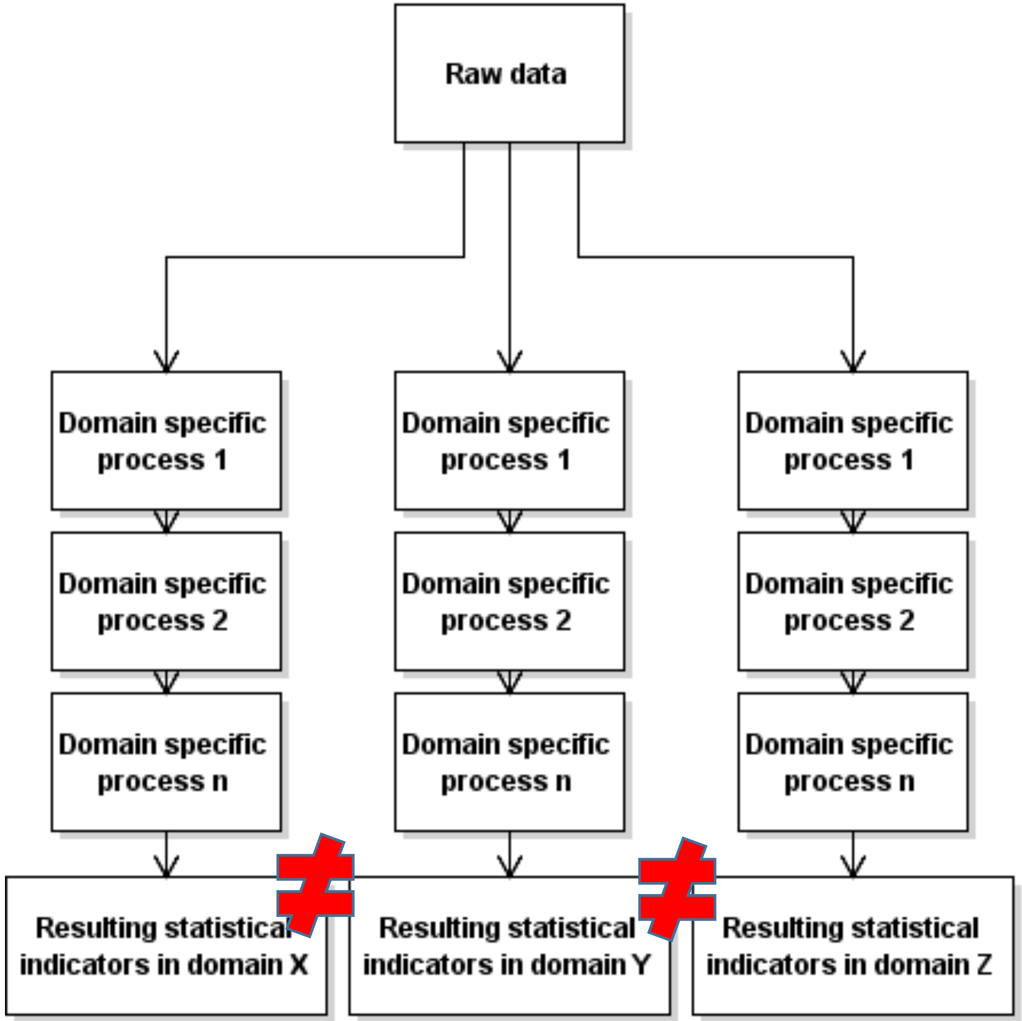
All trips in MPD ⟷ Tourism trips in MPD

TOTAL for 23 kabupatens

- Core processes -

There are some core processes that repeat and should be uniform across different uses of the data

Processing Data for Different Domains



Raw data

Domain specific process 1

Domain specific process 2

Domain specific process n

Domain specific process 1

Domain specific process 2

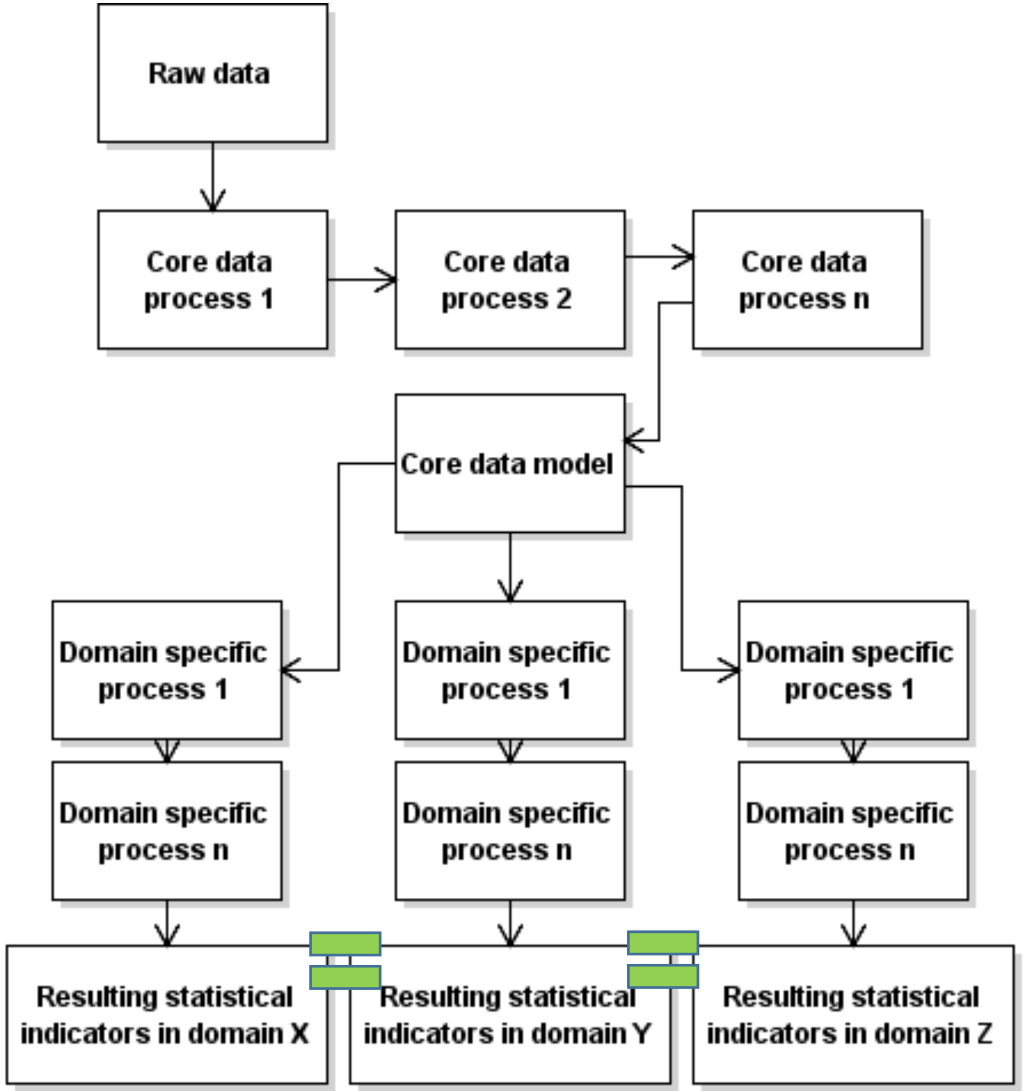
Domain specific process n

Domain specific process 1

Domain specific process 2

Domain specific process n

Resulting statistical indicators in domain X ≠ Resulting statistical indicators in domain Y ≠ Resulting statistical indicators in domain Z



Raw data

Core data process 1

Core data process 2

Core data process n

Core data model

Domain specific process 1

Domain specific process n

Domain specific process 1

Domain specific process n

Domain specific process 1

Domain specific process n

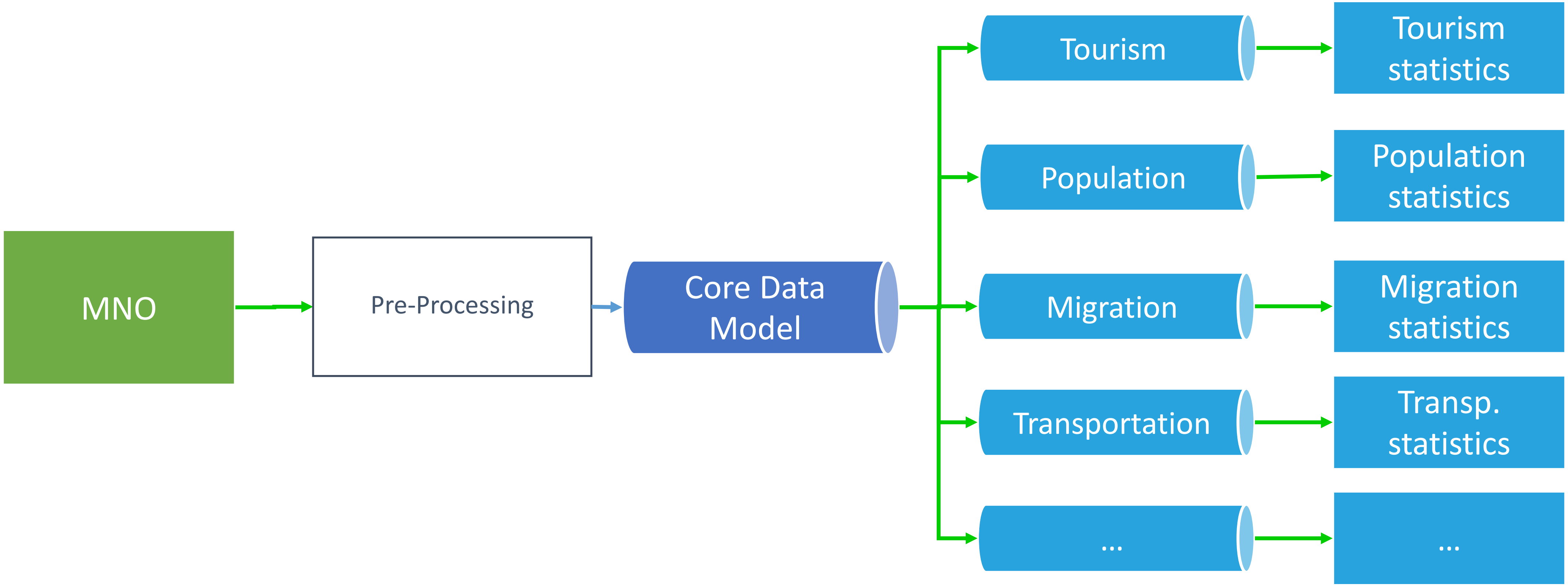
Resulting statistical indicators in domain X = Resulting statistical indicators in domain Y = Resulting statistical indicators in domain Z

Examples of core processes

1. Input data QA
2. Cleaning of noise
3. Trip generation
4. Home detection and usual environment

These steps are completed in a unified way for different domains

Core Data Model



- Quality Assurance never stops -

QA is a consistent part of every process step

Statistics quality frameworks

- UNECE suggested framework for the quality of big data
  - Covers the 3 phases of statistical production:
    - Input – data is acquired or in the process of being acquired
    - Throughput – data is transformed, analysed and manipulated
    - Output – the resulting statistics

Statistics quality frameworks

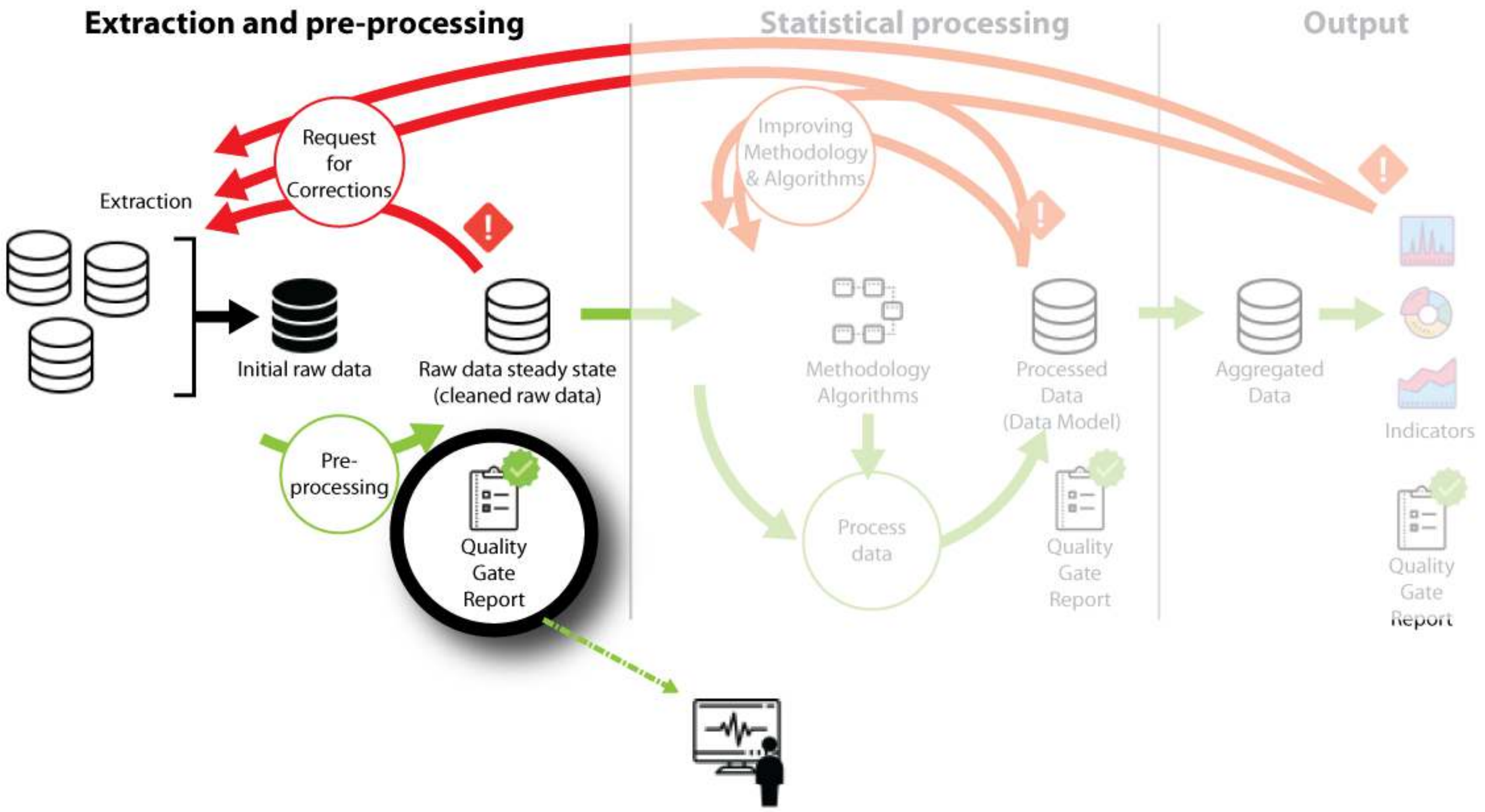
- UNECE suggested framework for the quality of big data
  - 3 hyperdimensions (objects which quality is assessed):
    - Source (type of data, characteristics of entity from which data is obtained, governance)
    - Metadata
    - Data

Quality Assurance Framework

Processing steps

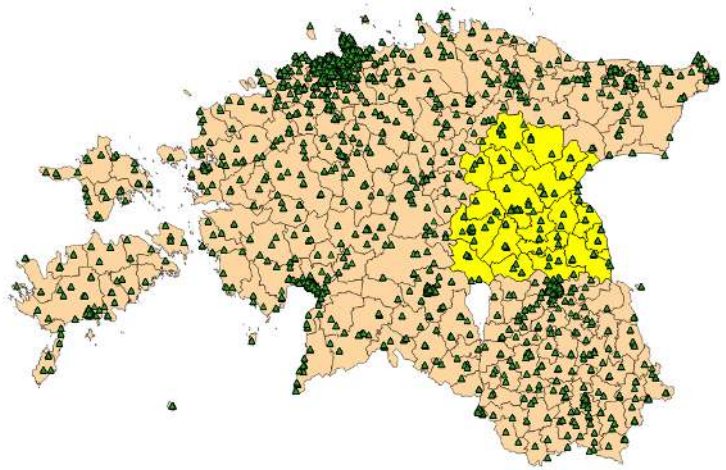
	Input	Throughput	Output
Source	Privacy and security		Confidentiality
Metadata	Log files Metadata Consistency …	System independence Quality gates Steady states	Accessibility and clarity Relevance
Data	Consistency Validity …		Coherence Consistency Validity …

Quality Gate 1 – Raw Data

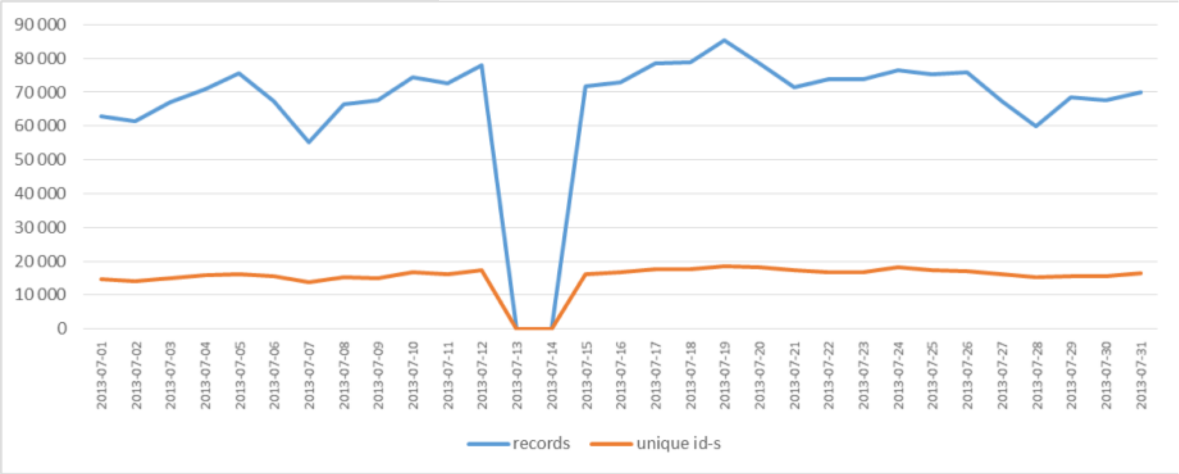


1

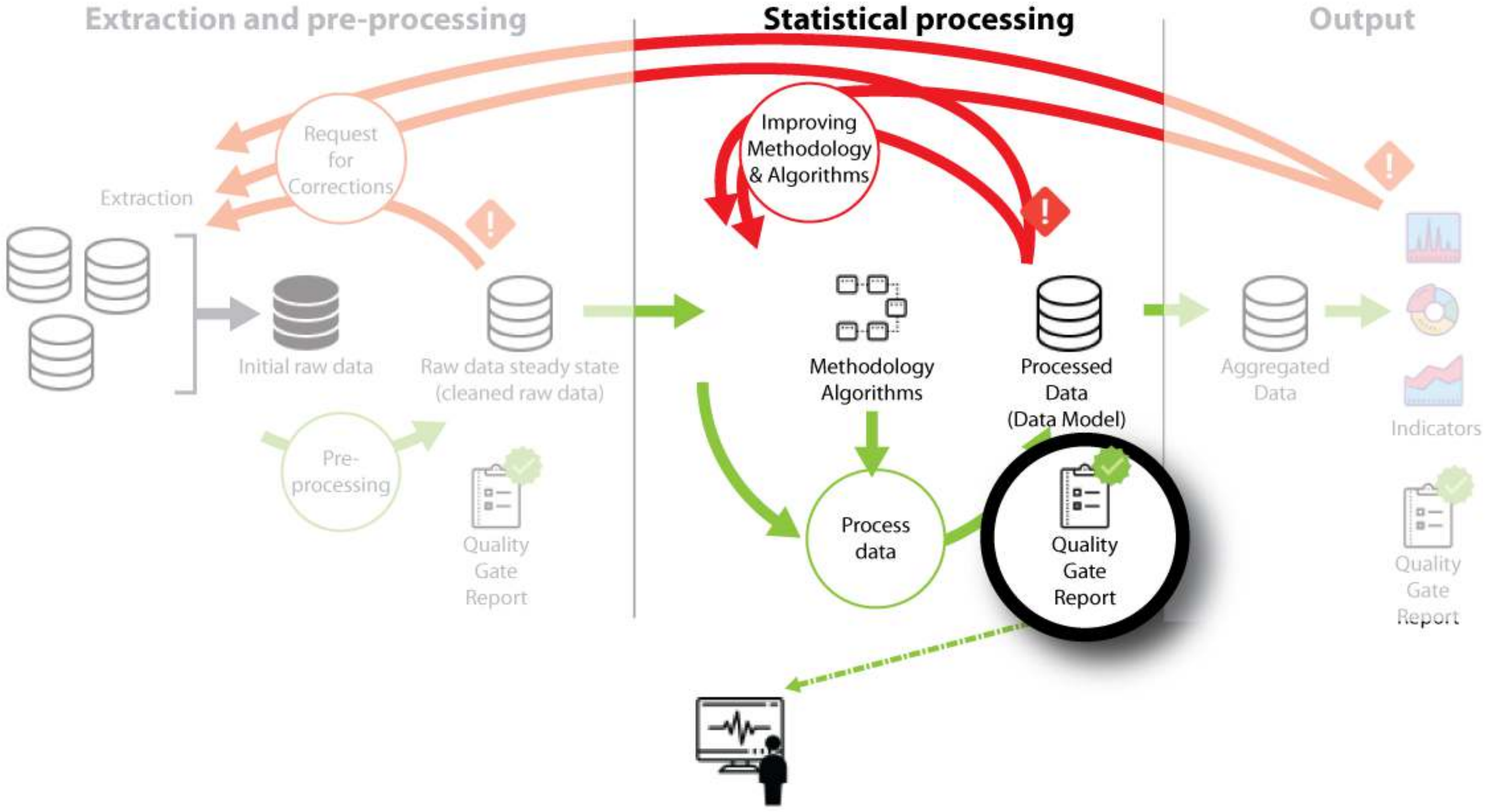
Common errors in raw data



- Wrong antenna coordinates or attributes
- Errors in antenna coordinates transformation
- Data gaps
- Missing data from some sub part of the system
- Time zone issues
- Incorrect format of timestamps
- Changes in continuity of the ID-s
- Duplicated records
- …



Quality Gate 2 – Modelled Data

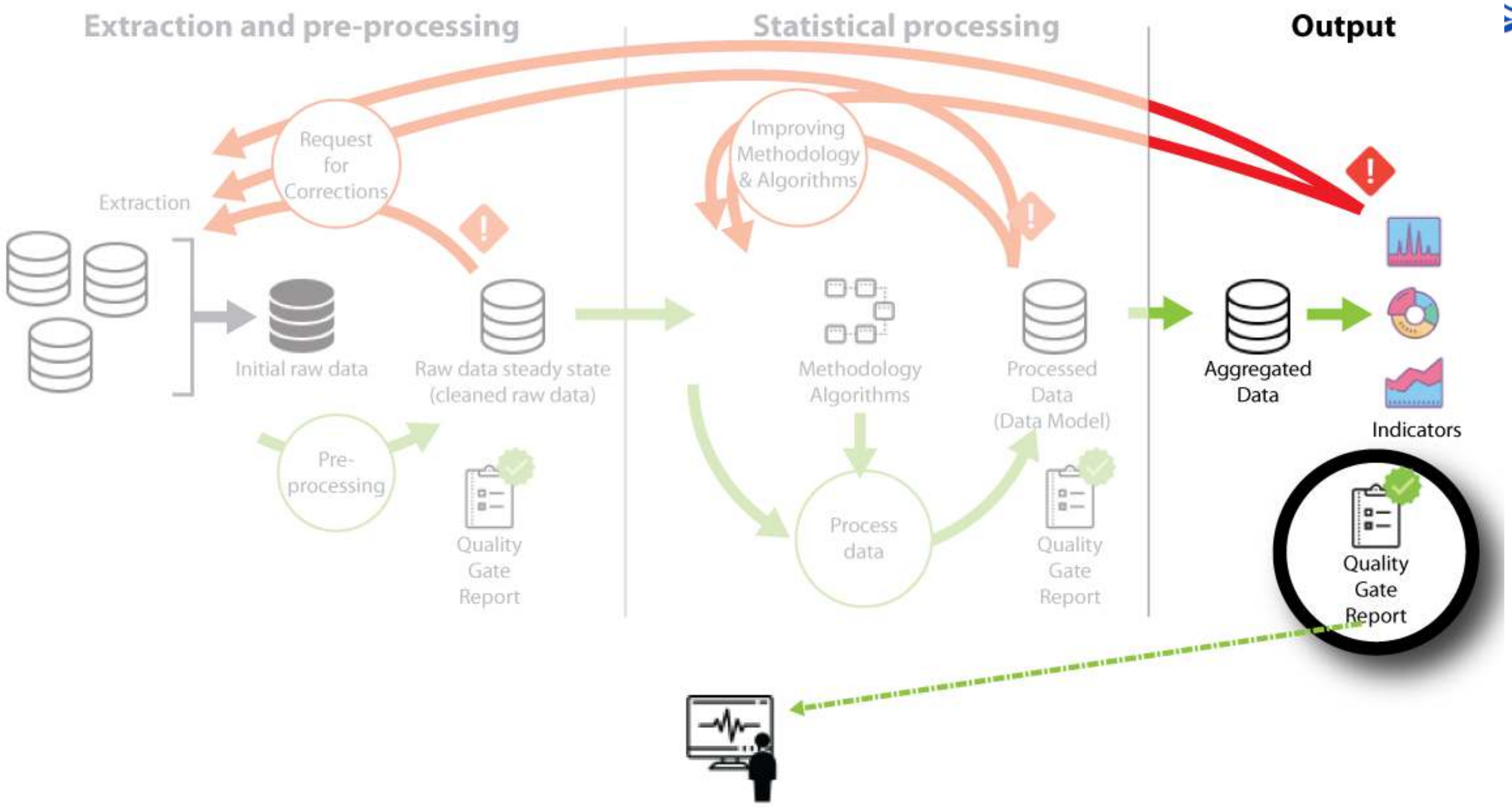


2

Common errors in processing

- Process produces an error
- Process does not finish
- Process ingests erroneous data
- Process overwrites critical data
- …

Quality Gate 3 – Output Data



3

Common errors in output data

If all processes run against correct methods and run correctly, output data should be sound. However,

- Low coherence to validation data
- Anomalies in the data
  - Peaks
  - Valleys
  - Gaps
- Trends that indicate a systematic change in underlying data
- New phenomena