Towards a Reference Framework for the processing of mobile network operator data for official statistics

Mihail Skaliotis & Fabio Ricciato
EUROSTAT Task Force on Big Data
Dubai, 21.10.2018

Motivations

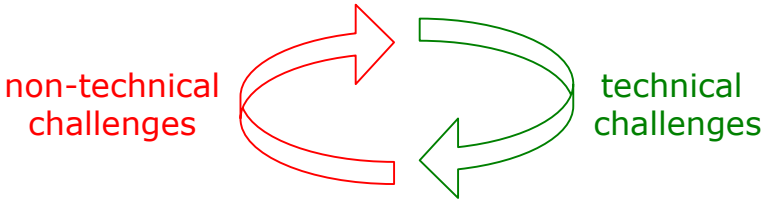
15+ years of research on case-study shows that MNO data have potential for Official Statistics ☺

but ☹

- *no unified/systematic methodology (clutter of ad-hoc methods for specific case-studies)*
- *poor reproducibilty, difficult portability*
- *rigidity, lack of evolvability*
- *some legal aspects still unclear (privacy & business confidentiality)*
- *no cross-domain analysis (multi-MNO)*

MNO : mobile network operators

Our Goals



Develop a unified methodological view:

Reference Methodological Framework [RMF] for processing MNO data for Official Statistics

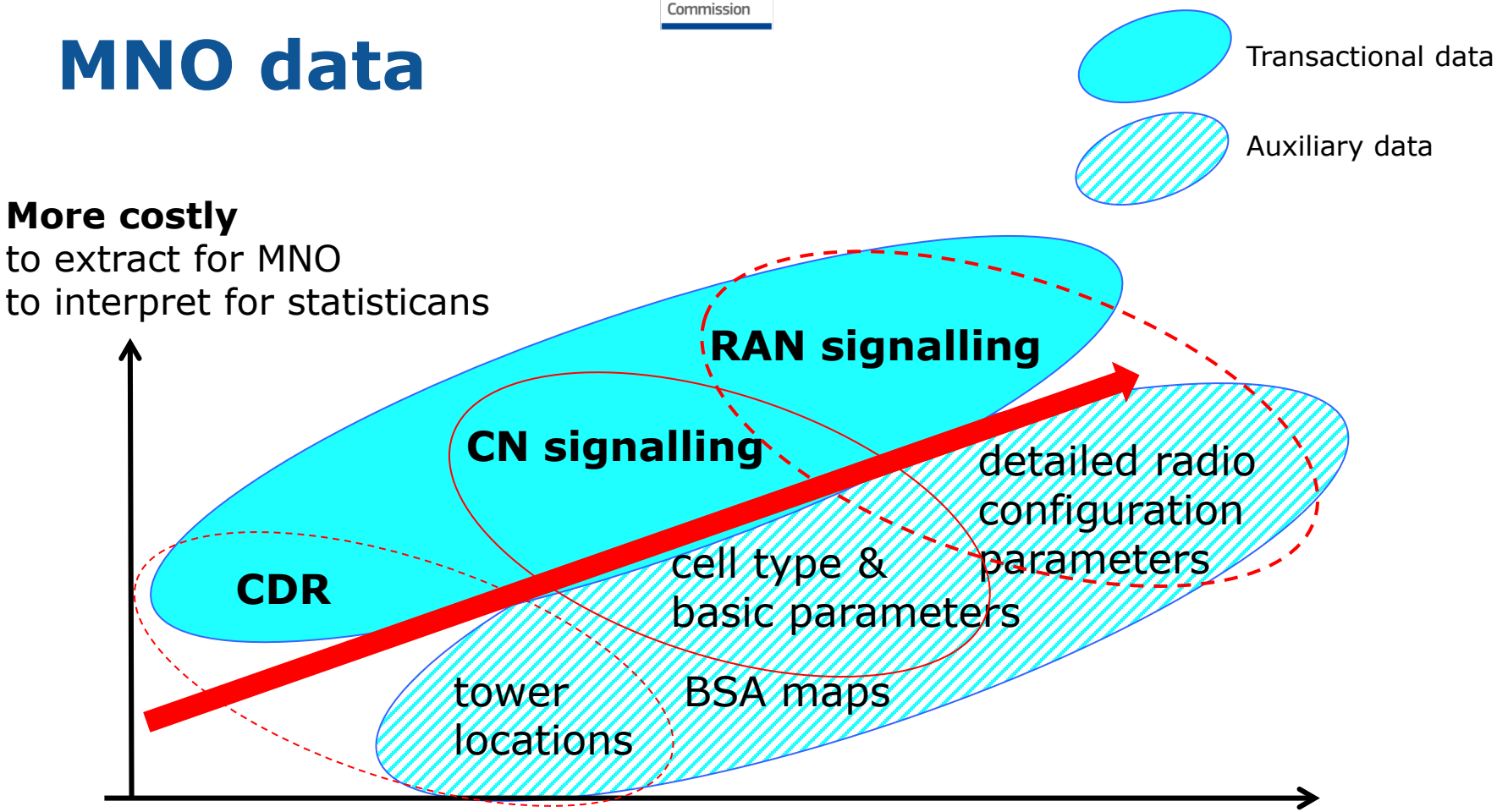
in order to:

- *facilitate interworking MNO-ESS at technical & organisational level*
- *ensure consistency, reproducibilty, **evolvability** and portability of processing methods (between MNOs and NSIs)*
- *provide concrete basis to clarify legal aspects (→ GDPR)*
- *enable multi-MNO analysis (fusion of data from different MNO)*

MNO: Mobile Network Operator

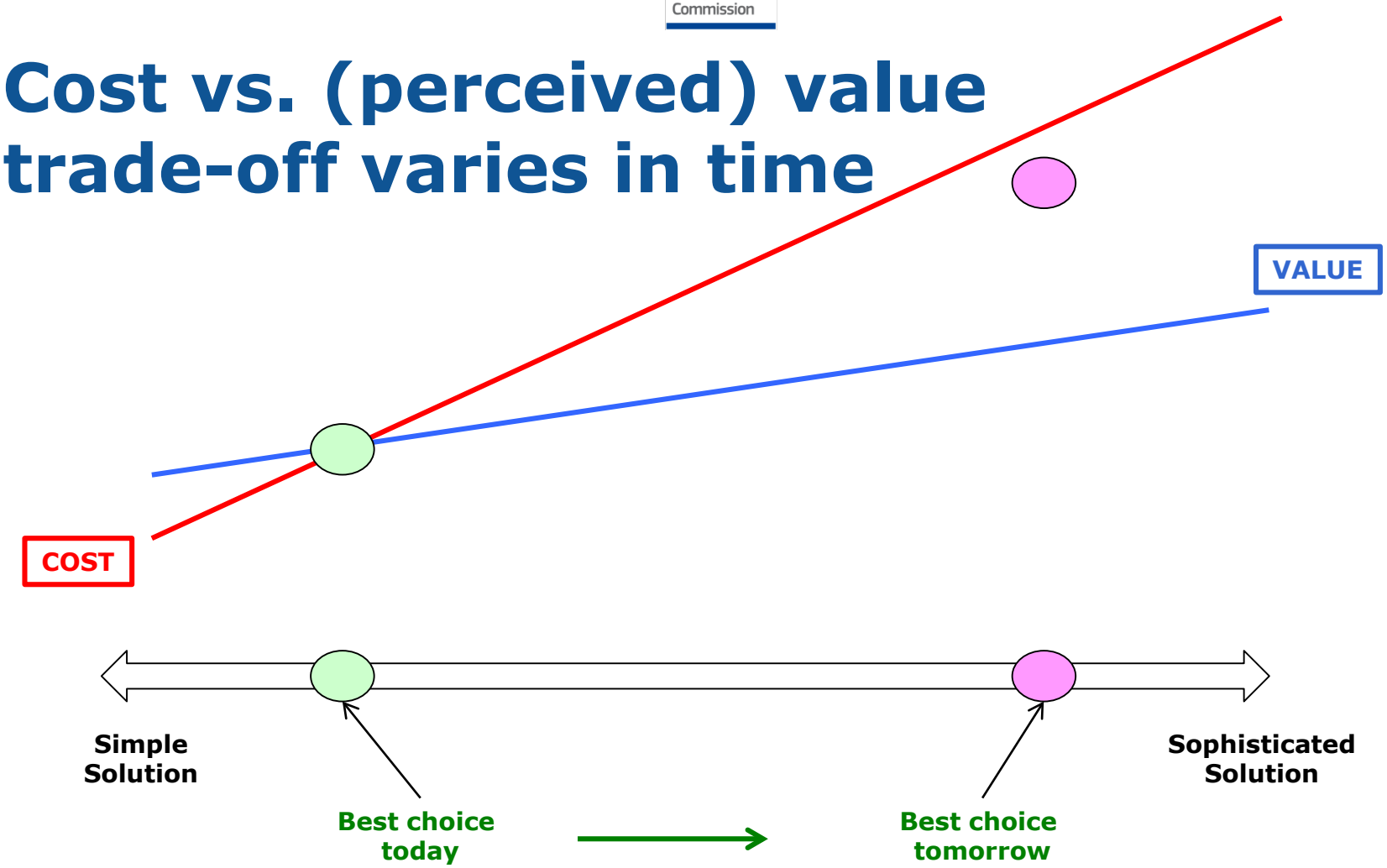
ESS: European Statistical System

MNO data

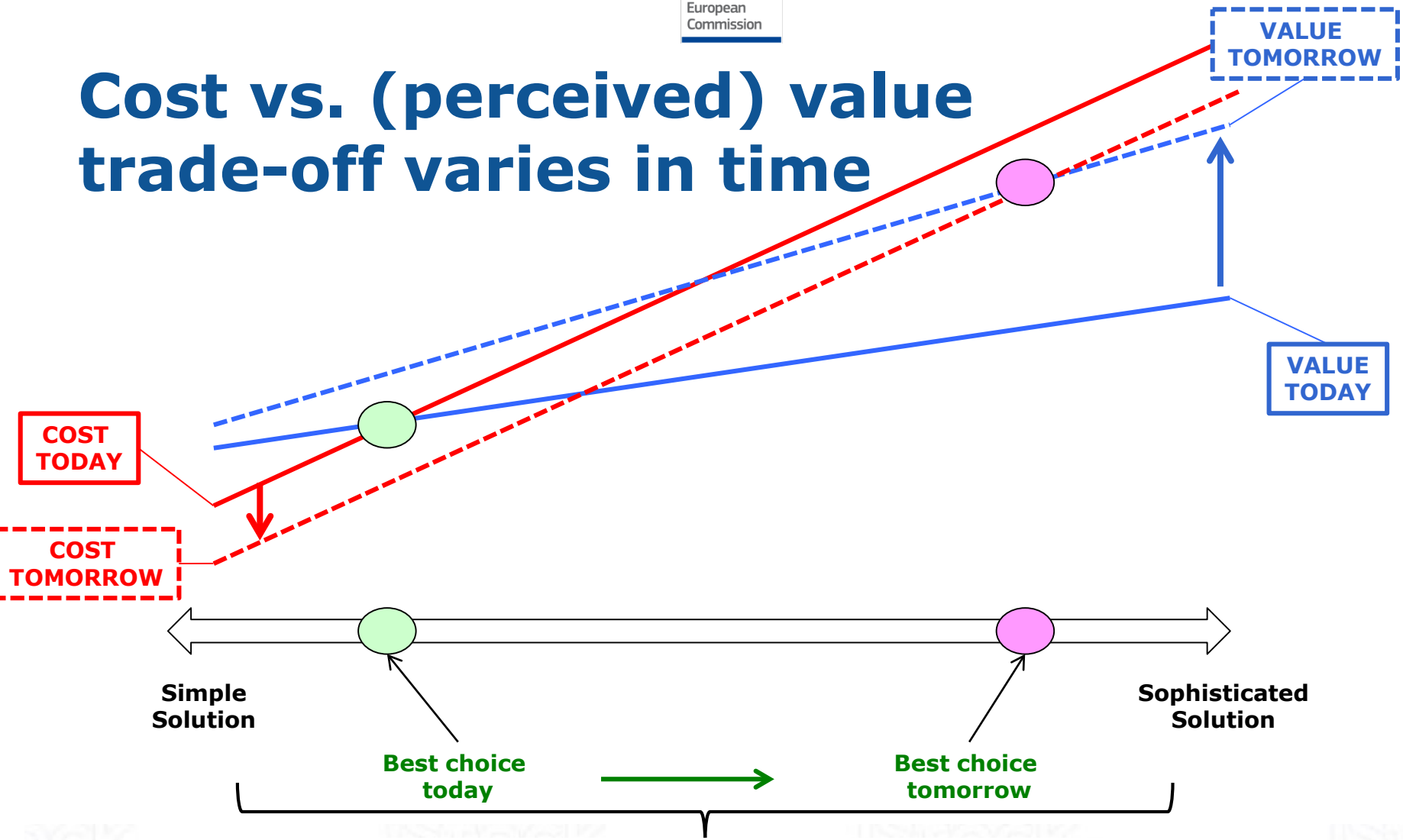


CN: Core Network
RAN: Radio Access Network
BSA: Best Service Area

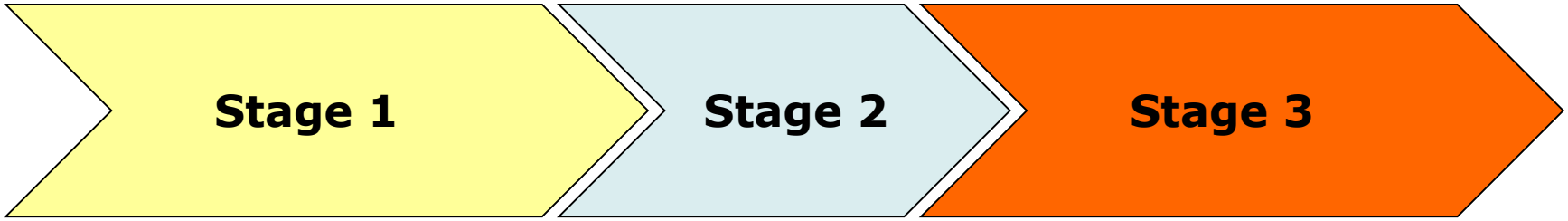
Cost vs. (perceived) value trade-off varies in time



Cost vs. (perceived) value trade-off varies in time



Ongoing work by EUROSTAT: staged approach



Single MNO

- definition of a reference layered architecture (hourglass model) and common data structures (C-layer)
- clarification of GDPR aspects

Multiple MNOs with *output* data fusion on NSI (silos model)

- testing and possible refinement of reference layered architecture across heterogeneous network operation settings

Multiple MNOs with *input* data fusion (via SMPC)

- definition of reference architecture for Secure Multi-Pparty Computation
- clarification of GDPR aspects related to SMPC

SMPC: Secure Muti-Party Computation

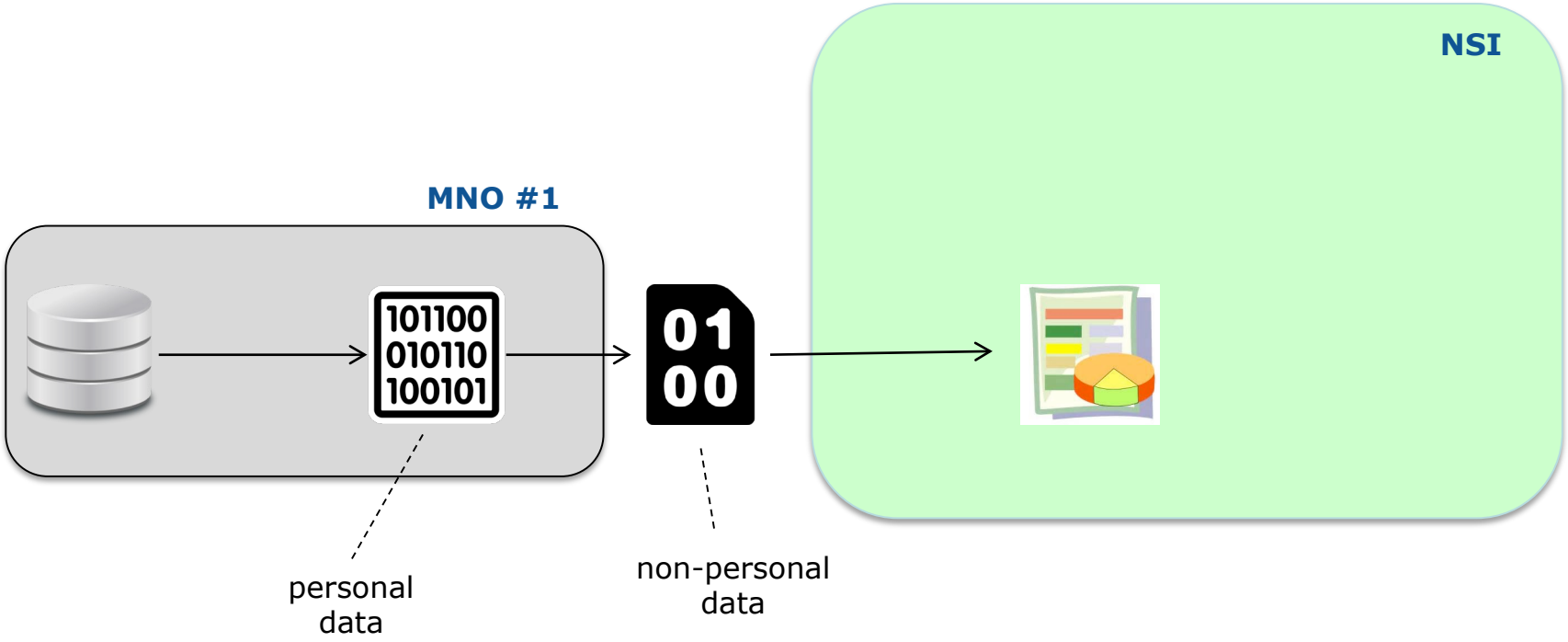
Stage 1 scenario

Raw micro-data (D-layer)

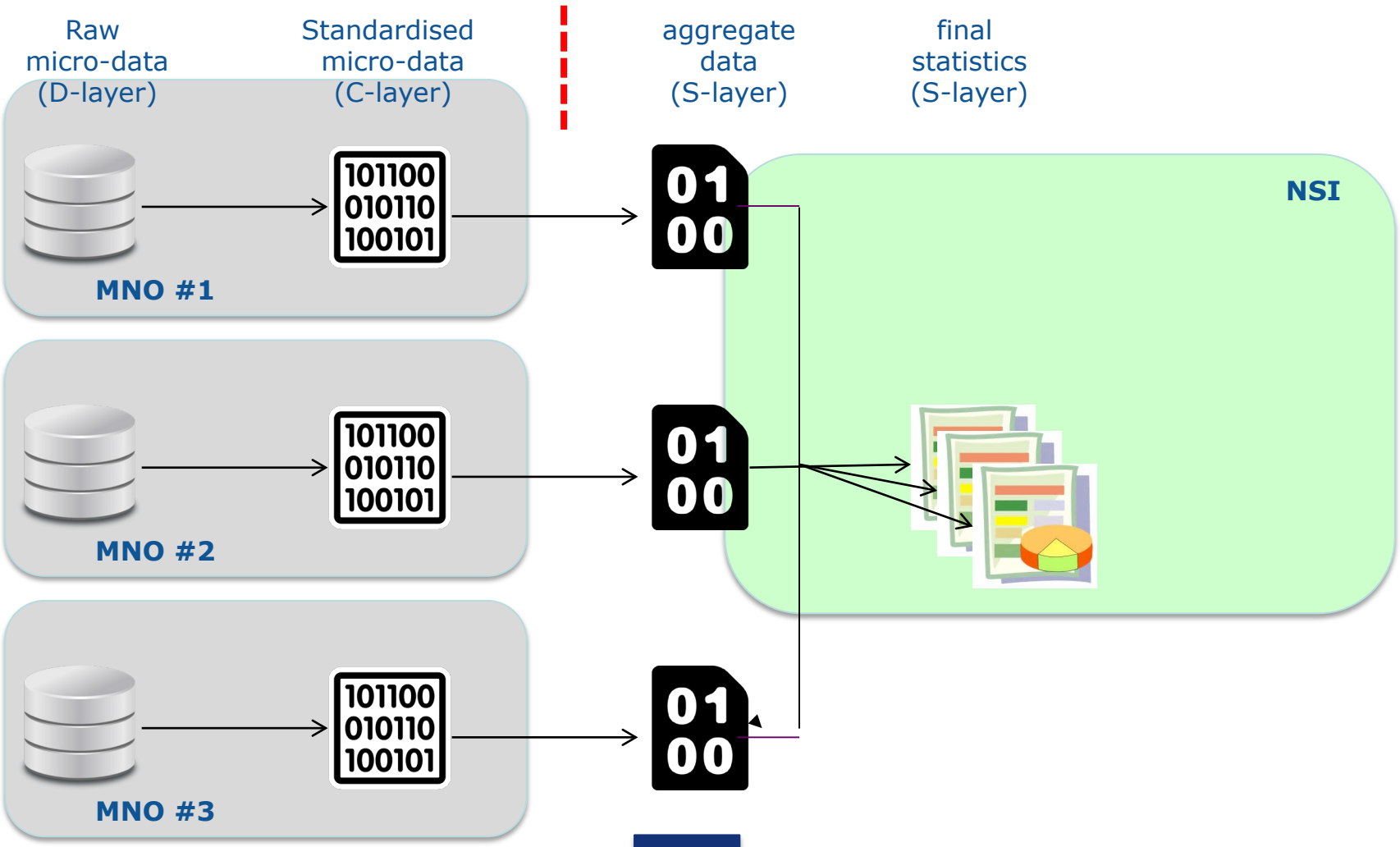
Standardised micro-data (C-layer)

aggregate data (S-layer)

final statistics (S-layer)

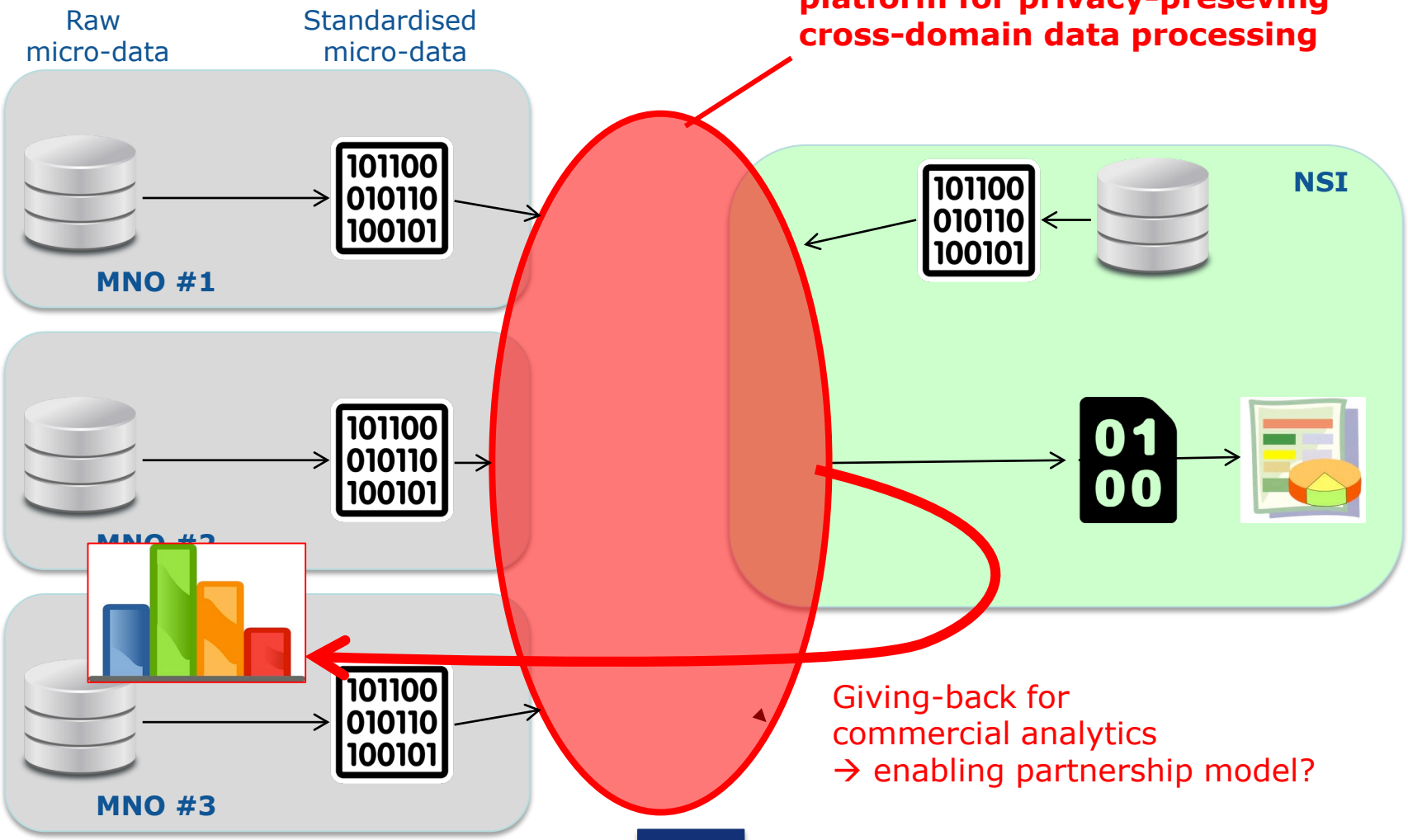


Stage 2 scenario



Stage 3 scenario

Secure Multiparty Computation (SMPC) platform for privacy-preseving cross-domain data processing

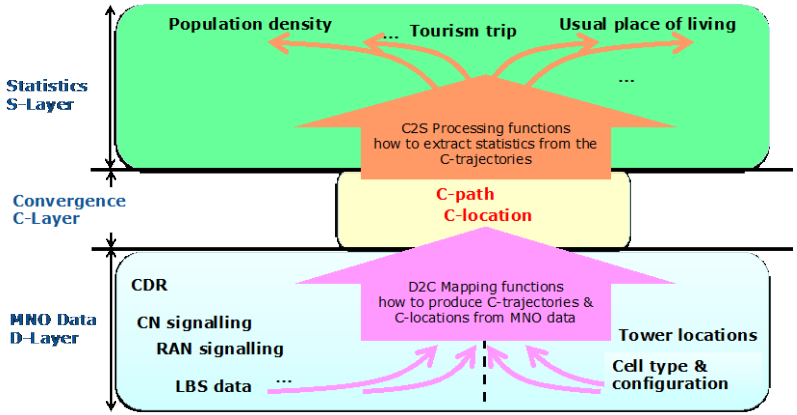


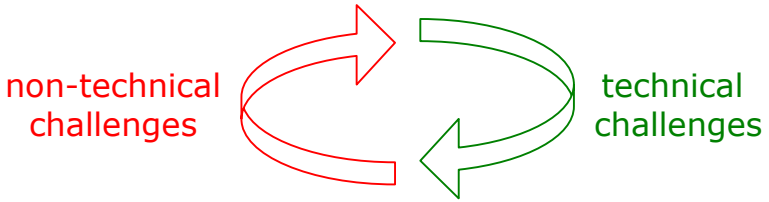
Stage 1 goals

- *Define the Reference Methodological Framework (RMF) for a single MNO data stream*

- *Proof-of-concept application of RMF on selected use-case*
  - population density (ongoing work 2018)
  - tourism (next year)

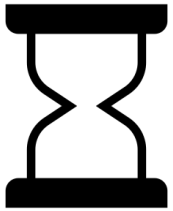
- *Clarify GDPR aspects*
  - started dialogue with European Data Protection Supervisor





- ***Collaboration EUROSTAT-Proximus***
- ***Dedicated WP in future ESSnet on Trusted Smart Statistics***

Hourglass model



Statistics
S-Layer

Heterogeneity of applications & use-cases
Diversity of statistical definitions
Complexity of statistical objects
Multiple NSIs

Domain of Expertise
Statisticians, NSI

Convergence
C-layer

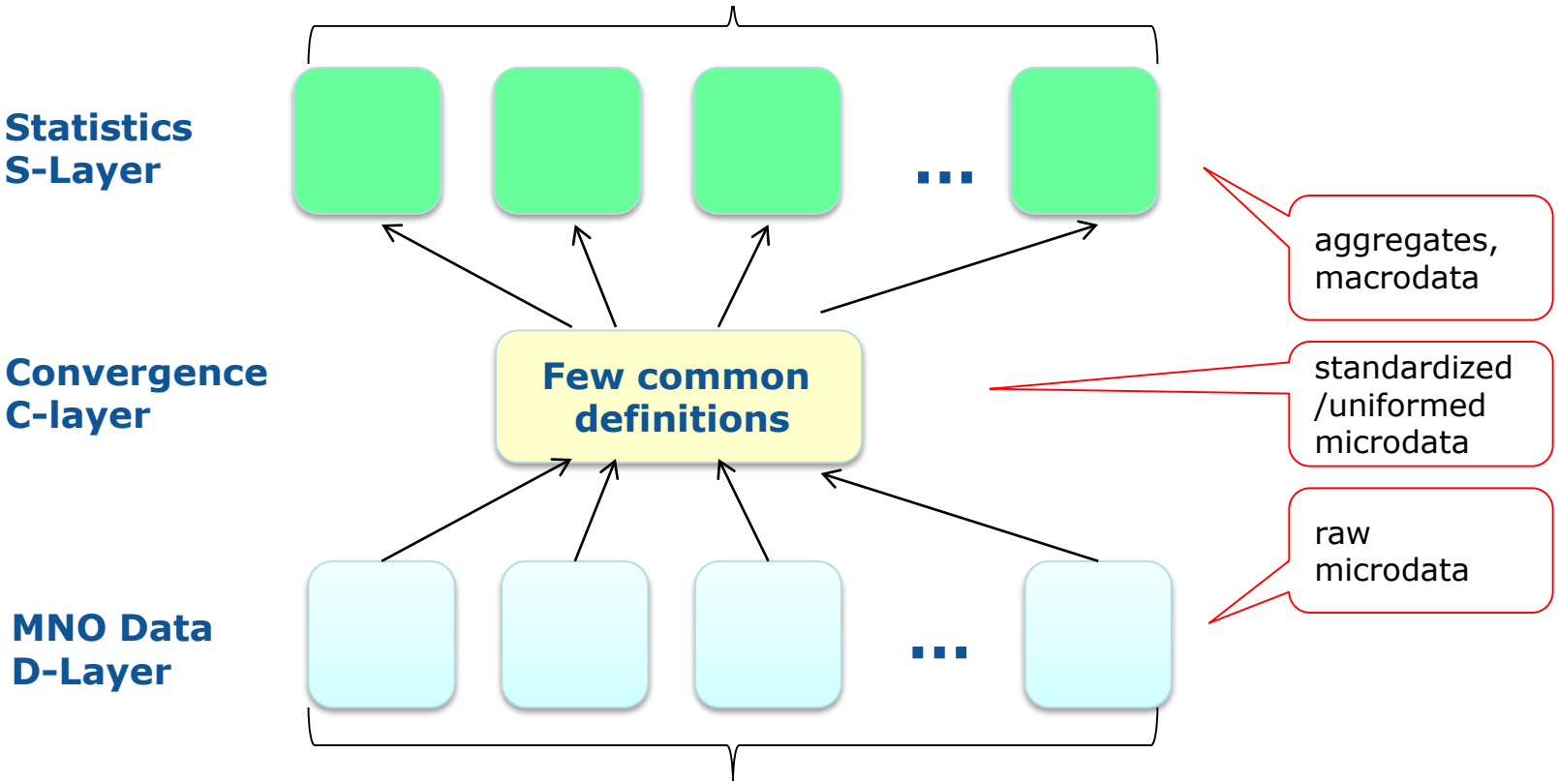
Few common definitions

MNO Data
D-Layer

Data Heterogeneity
Diversity of data collection methods
Complexity of data semantics
Multiple MNOs

Domain of Expertise
Telco Engineers, MNO

Multiple data consumers: ESTAT, NSI#1, NSI#2…
Different subject matter experts & use-cases: tourism, population, transport, …



Multiple data sources: MNO#1, MNO#2…
Different data types: CDR, signalling data, RAN data, LBS, …

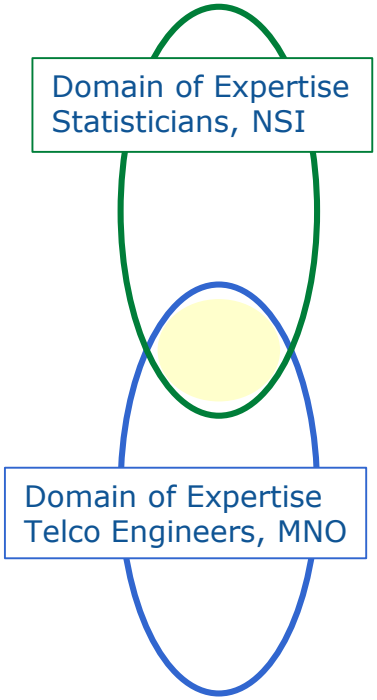
Benefits of layering

Decouples the complexity & heterogeneity of the two domains

- Hides complexity & heterogeneity of MNO data to statisticians
- Hides complexity & heterogeneity of statistical concepts to MNO engineers

*Decoupling allows for independent **development, adoption** & **evolution** at each domain*

Requirements for C-layer



Parsimony, Clarity

- Few definitions that are understood and accepted by experts of both domains

Feasibility

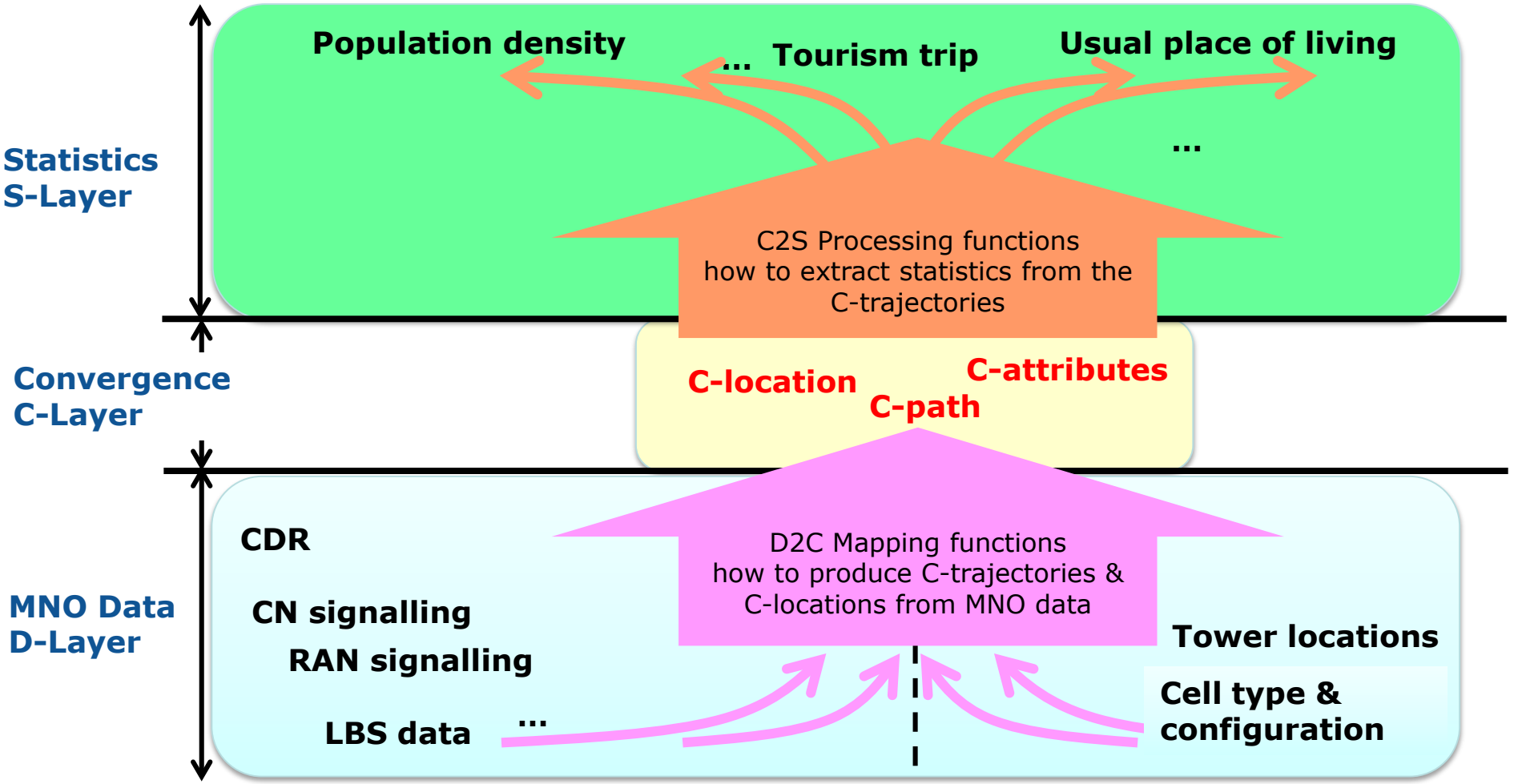
- C-layer structures must represent what can be realistically obtained from underlying MNO data

Sufficiency

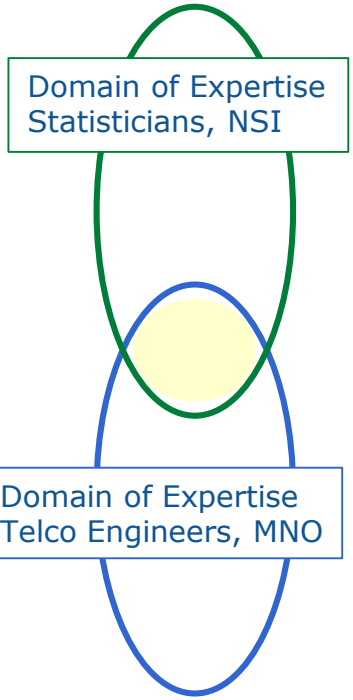
- C-layer structures must be informative for the use-cases above

Generality

- Not tailored to specific MNO dataset and/or use-case here&now: basis for long-term adoption & evolution


Statistics
S-Layer
Population density
... Tourism trip
Usual place of living
...
C2S Processing functions
how to extract statistics from the
C-trajectories
Convergence
C-Layer
C-location
C-attributes
C-path
MNO Data
D-Layer
CDR
CN signalling
RAN signalling
LBS data
...
D2C Mapping functions
how to produce C-trajectories &
C-locations from MNO data
Tower locations
Cell type &
configuration

The C-layer is …

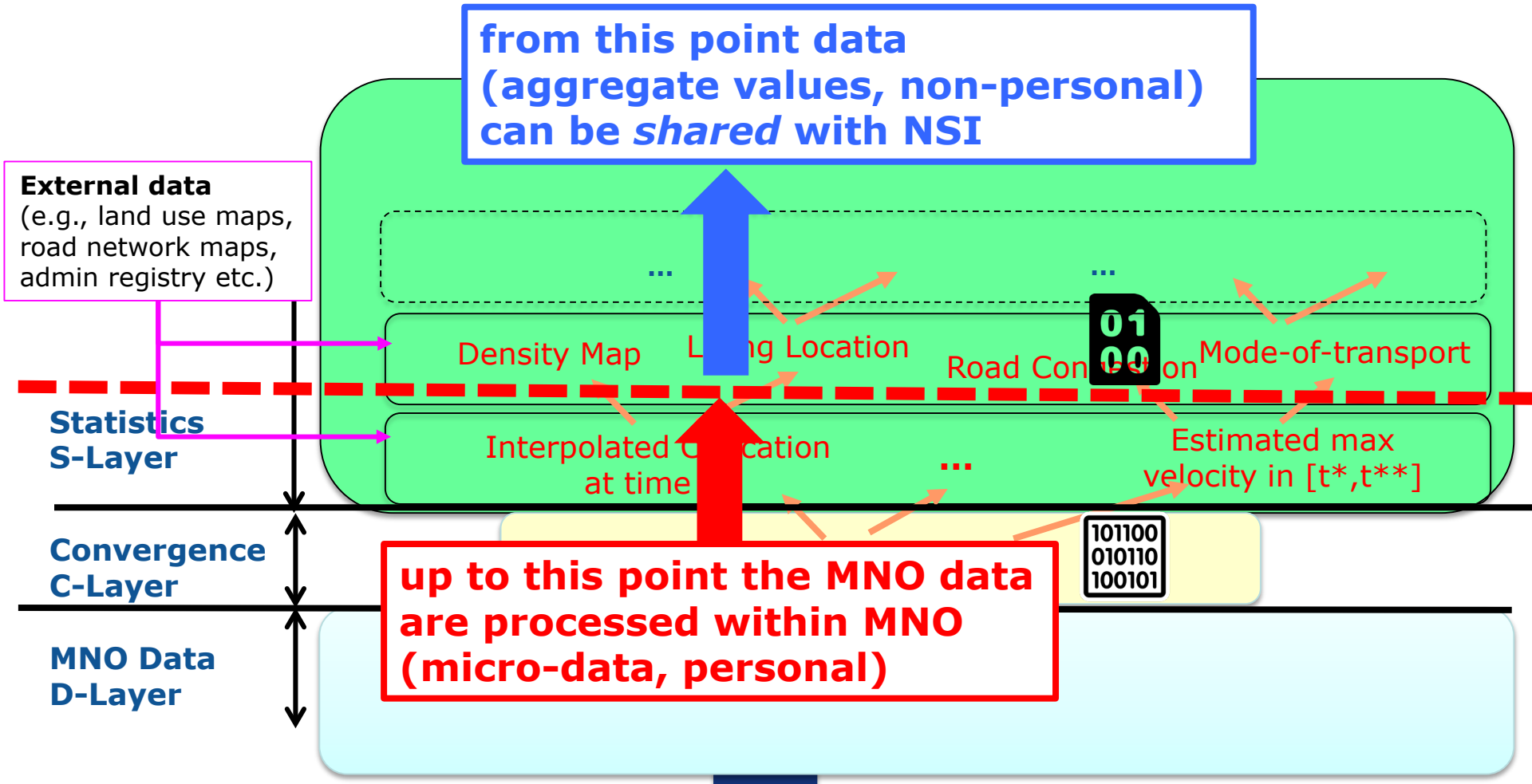


The C-layer

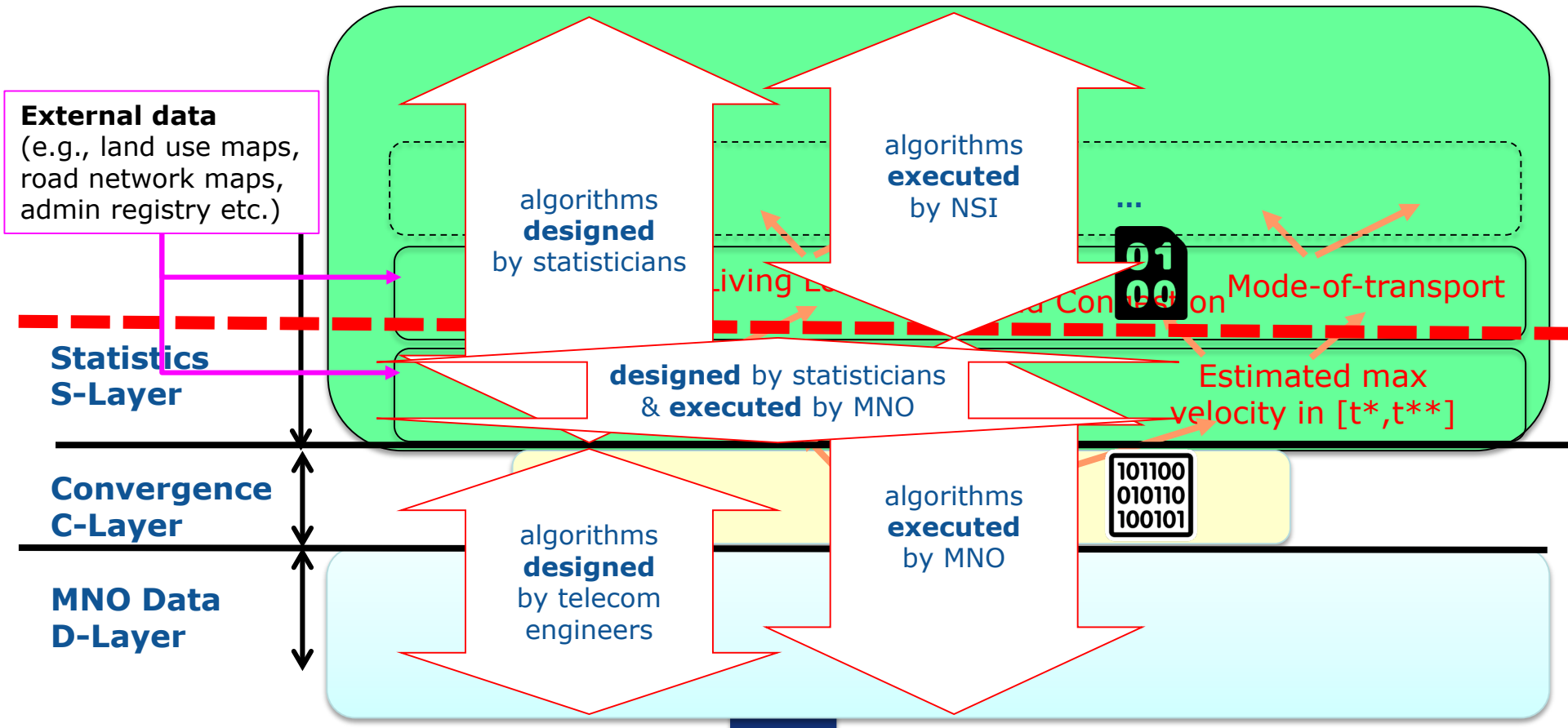
is an abstract "interface" between knowledge domains
(statisticians – telecom engineers)
*it is relevant for the **design** of processing method*
(algorithm)

it is NOT a physical interface for data export!
(such interface, relevant for the execution of the computation process, is logically placed within the S-layer)

NB: the MNO-to-NSI exporting interface is within the S-layer!



algorithm design vs execution



Stage 1 scenario

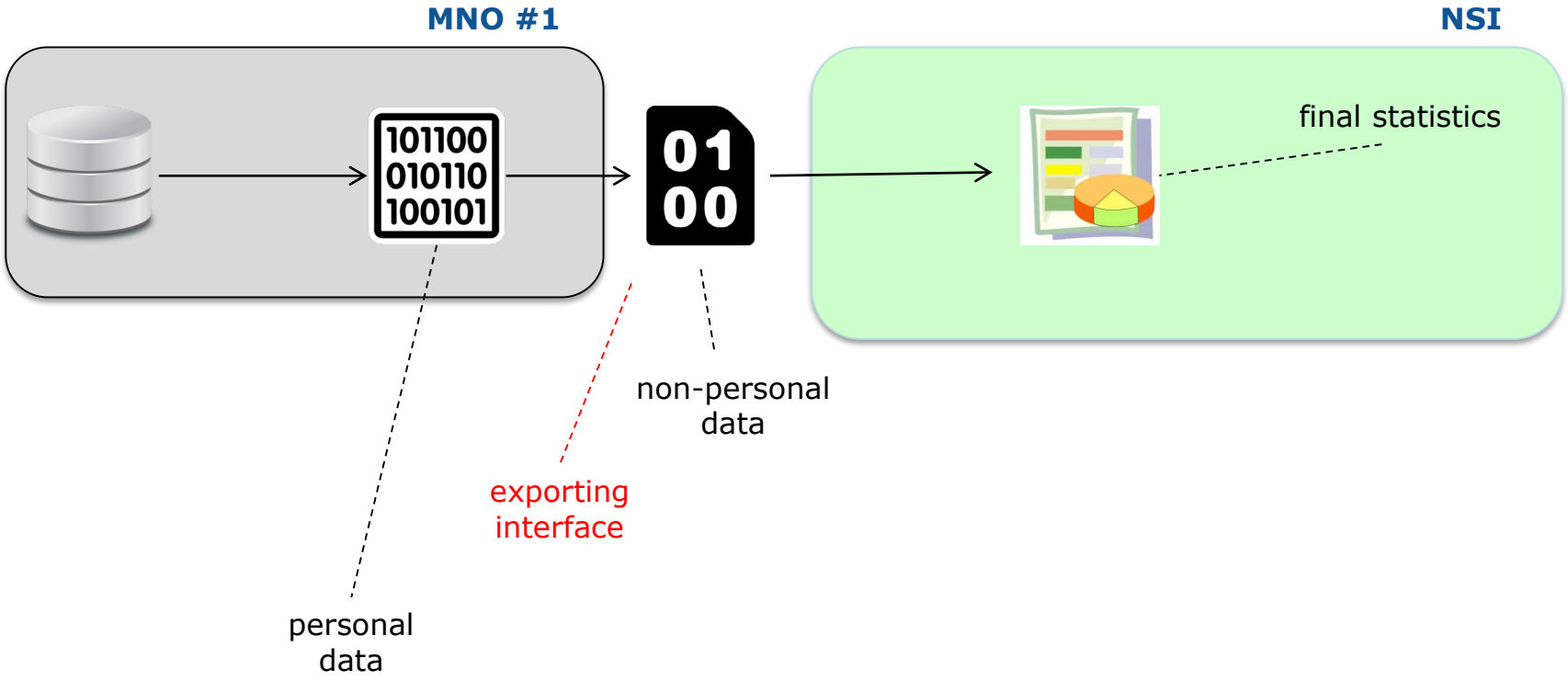
Stage 1

Raw micro-data (D-layer)

Standardised micro-data (C-layer)

aggregate data (S-layer)

final statistics (S-layer)



Stage 1 scenario

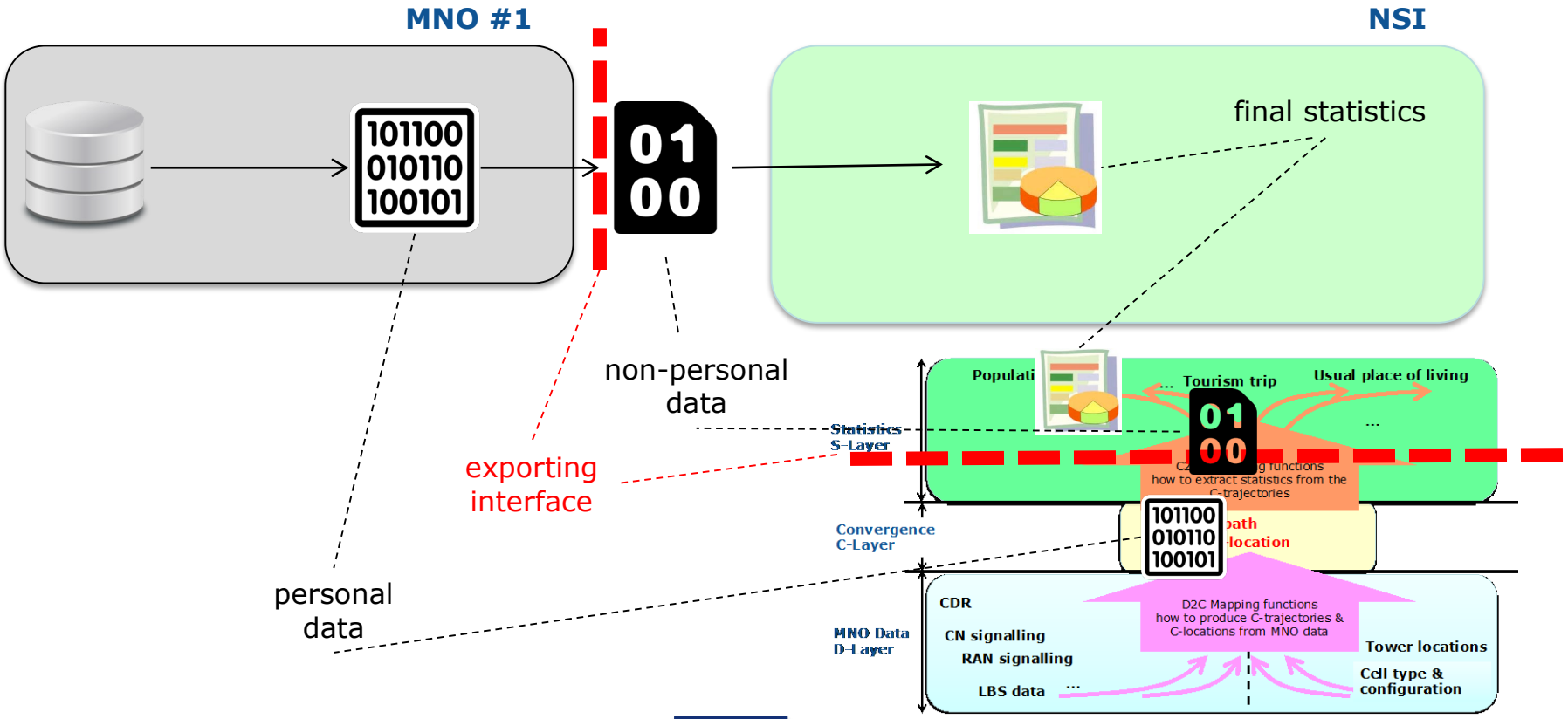
Stage 1

Raw micro-data (D-layer)

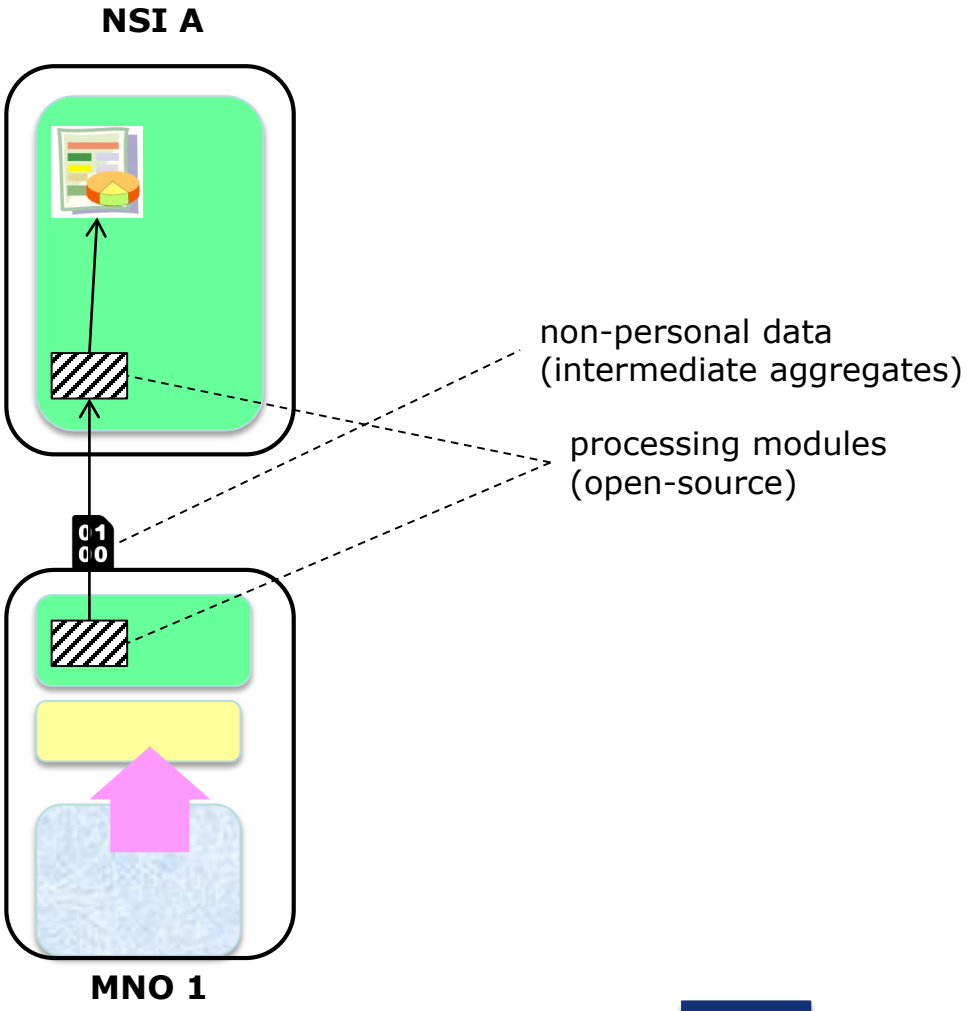
Standardised micro-data (C-layer)

aggregate data (S-layer)

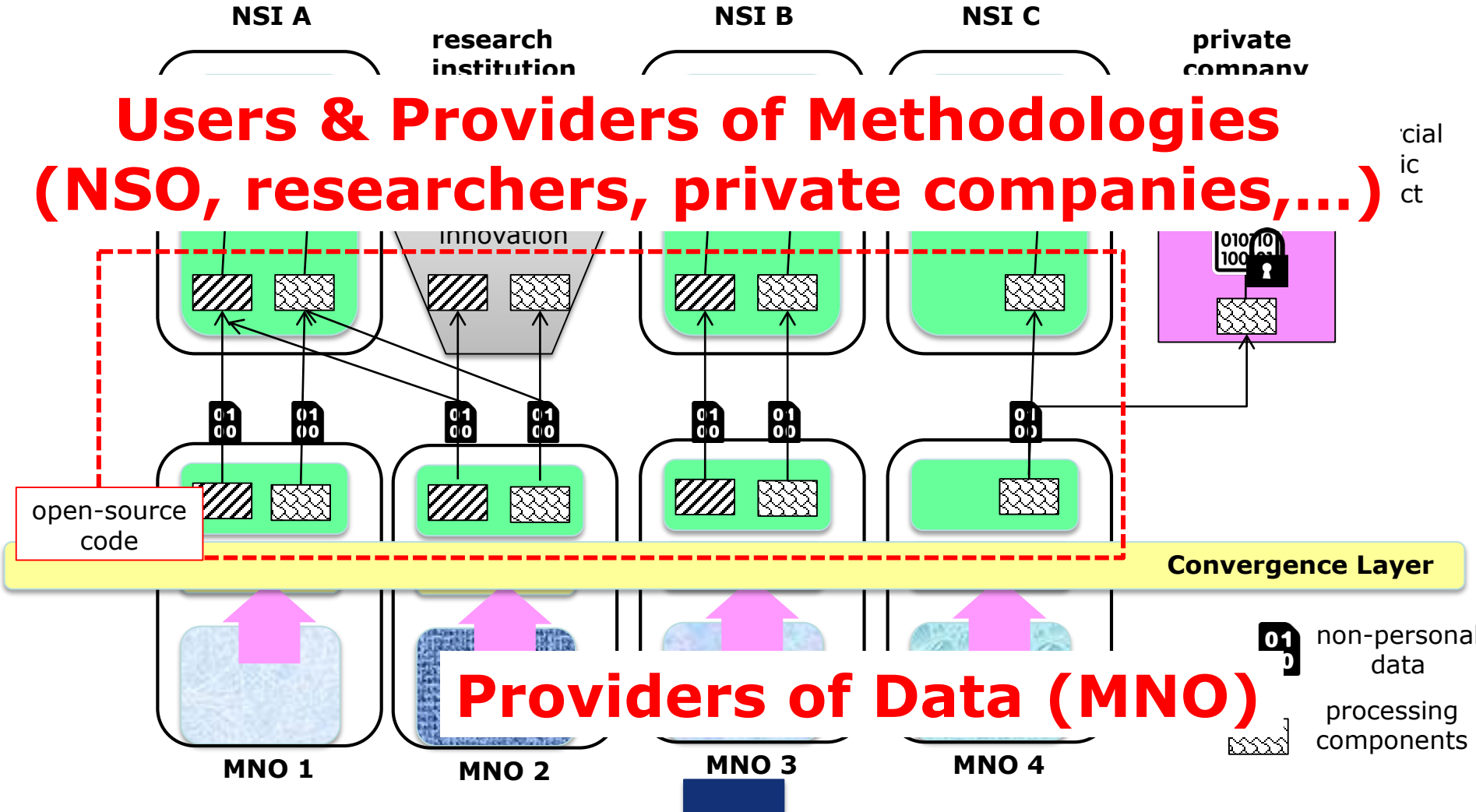
final statistics (S-layer)



C-layer as a common substratum for MNO data users



C-layer as a common substratum for MNO data users



Thanks for your attention

For follow-up:

fabio.ricciato@ec.europa.eu