Overview of current activities in MNO data at EUROSTAT

Fabio Ricciato, EUROSTAT

Project Meeting on Measuring Human Mobility

27 - 29 March 2019, Tbilisi, Georgia

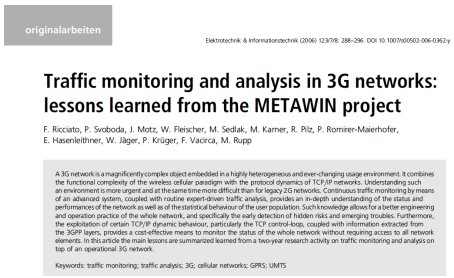
About EUROSTAT and ESS

- *Eurostat is the statistical office of the European Union, and is part of the European Commission*

- *The European Statistical System (ESS) is the partnership between **Eurostat**, the national statistical institutes (**NSIs**) and other national authorities responsible in each Member State for the development, production and dissemination of European statistics.*

About me

- *MS'99 Electrical Engineering, PhD'03 in Telecommunications, from Univ. La Sapienza, Rome*
- *2004-2008 Senior researcher and project manager in FTW, Vienna*
  - METAWIN & DARWIN projects on 3G data traffic monitoring
- *2007-2013 Assistant Professor, Univ. of Salento, Lecce, Italy*
  - Teaching Telecommunication Systems (2G/3G/4G networks)
- *2013-2014 Head of Business Unit at AIT, Vienna*
  - JRC study on density estimation from mobile network data
- *2015-2017 Professor at Univ. of Ljubljana, Slovenia*
- *Since January 2018 – Statistical offices in EUROSTAT*
  Unit B1 "Innovation; Methodologies in Official Statistics"


originalarbeiten

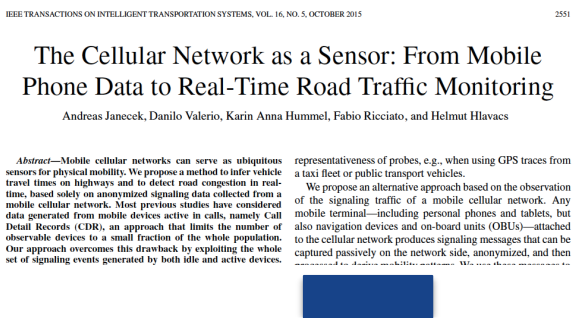
Elektrotechnik & Informationstechnik (2006) 123/7/8: 288–296. DOI 10.1007/s00502-006-0362-y

Traffic monitoring and analysis in 3G networks: lessons learned from the METAWIN project

F. Ricciato, P. Svoboda, J. Motz, W. Fleischer, M. Sedlak, M. Karner, R. Pilz, P. Romirer-Maierhofer, E. Hasenleithner, W. Jäger, P. Krüger, F. Vacirca, M. Rupp

A 3G network is a magnificently complex object embedded in a highly heterogeneous and ever-changing usage environment. It combines the functional complexity of the wireless cellular paradigm with the protocol dynamics of TCP/IP networks. Understanding such an environment is more urgent and at the same time more difficult than for legacy 2G networks. Continuous traffic monitoring by means of an advanced system, coupled with routine expert-driven traffic analysis, provides an in-depth understanding of the status and performances of the network as well as of the statistical behaviour of the user population. Such knowledge allows for a better engineering and operation practice of the whole network, and specifically the early detection of hidden risks and emerging troubles. Furthermore, the exploitation of certain TCP/IP dynamic behaviour, particularly the TCP control-loop, coupled with information extracted from the 3GPP layers, provides a cost-effective means to monitor the status of the whole network without requiring access to all network elements. In this article the main lessons are summarized learned from a two-year research activity on traffic monitoring and analysis on top of an operational 3G network.

Keywords: traffic monitoring; traffic analysis; 3G; cellular networks; GPRS; UMTS


IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 16, NO. 5, OCTOBER 2015 2551

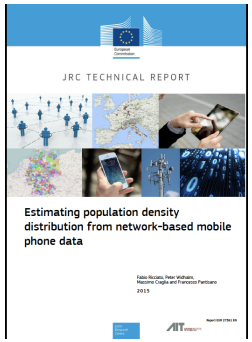
The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring

Andreas Janecek, Danilo Valerio, Karin Anna Hummel, Fabio Ricciato, and Helmut Hlavacs

***Abstract*—Mobile cellular networks can serve as ubiquitous sensors for physical mobility. We propose a method to infer vehicle travel times on highways and to detect road congestion in real-time, based solely on anonymized signaling data collected from a mobile cellular network. Most previous studies have considered data generated from mobile devices active in calls, namely Call Detail Records (CDR), an approach that limits the number of observable devices to a small fraction of the whole population. Our approach overcomes this drawback by exploiting the whole set of signaling events generated by both idle and active devices.**

representativeness of probes, e.g., when using GPS traces from a taxi fleet or public transport vehicles.

We propose an alternative approach based on the observation of the signaling traffic of a mobile cellular network. Any mobile terminal—including personal phones and tablets, but also navigation devices and on-board units (OBUs)—attached to the cellular network produces signaling messages that can be captured passively on the network side, anonymized, and then


JRC TECHNICAL REPORT

Estimating population density distribution from network-based mobile phone data

Fabio Ricciato, Peter Widhalm, Massimo Craglia and Francesco Pantisano

2015

Current activities on MNO data in Eurostat

- *Developing Reference Methodological Framework (RFM) for using MNO data for Official Statistics*
  - focus on presence & mobility patterns
  - CDR and signalling data

- *Developing and comparing different methodological variants for the density estimation problem*

- *Exploiting Secure Multi-party Computation (SMC) for multi-MNO data fusion*
  - capacity building, technical and legal aspects

Ongoing Collaborations

- *Cooperation with MNO*
  - proximus (belgium)

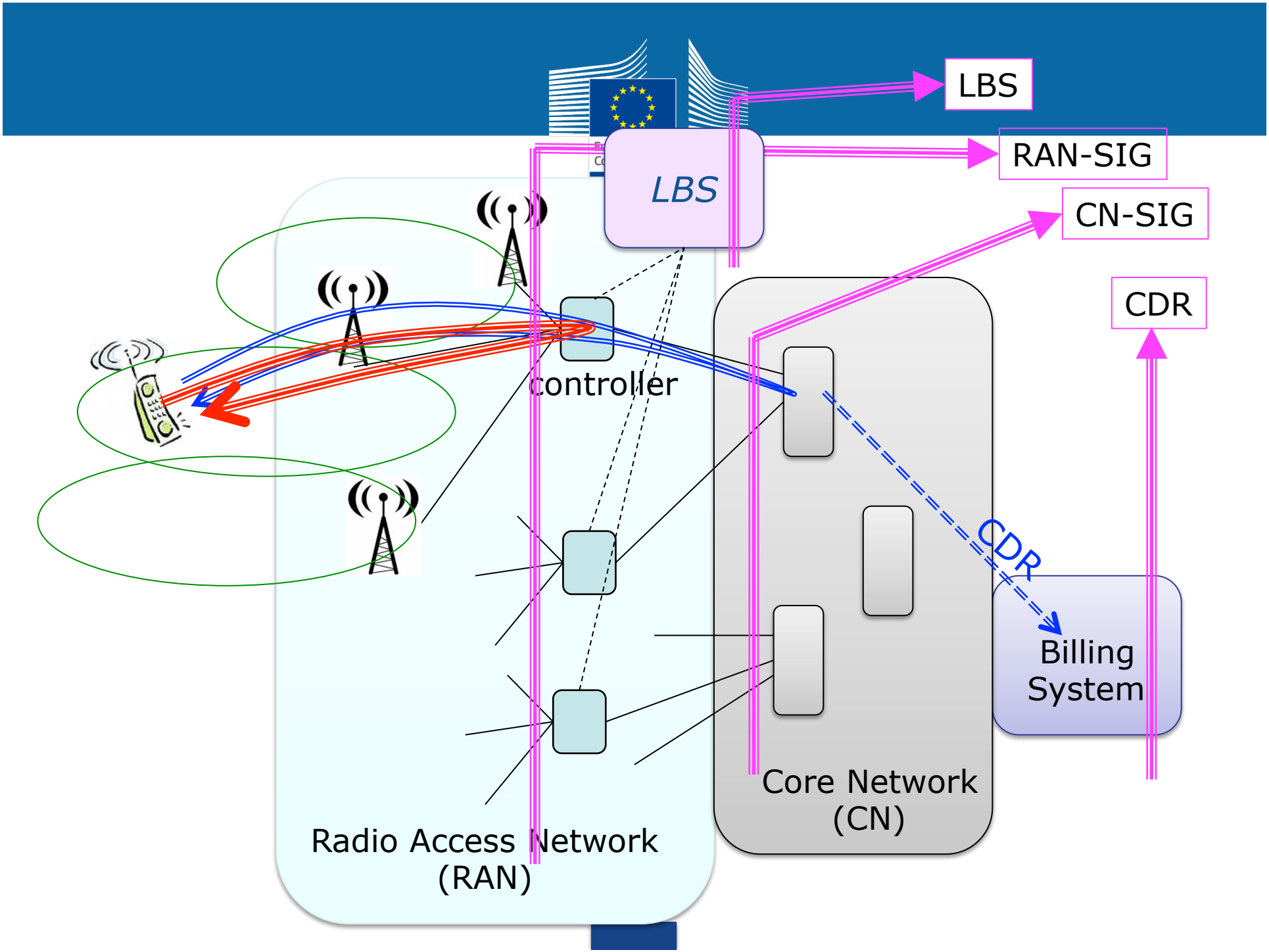
- *Cooperation with NSI*
  in the framework of ESSnet on Big Data
  - Italy, Netherlands, France, Spain,…

Current activities on MNO data in Eurostat

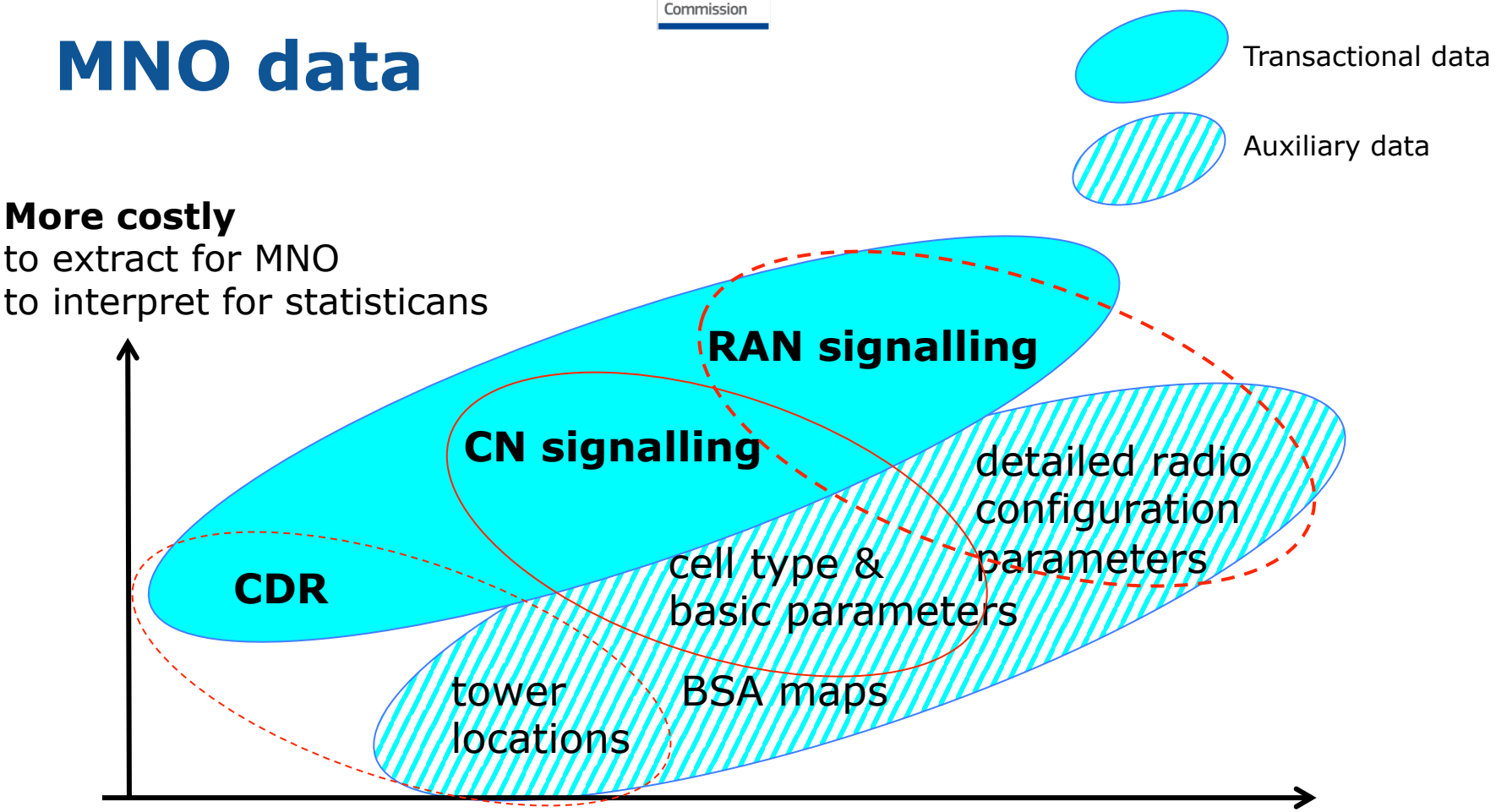
- *Developing Reference Methodological Framework (RFM) for using MNO data for Official Statistics*
  - focus on presence & mobility patterns
  - CDR and signalling data

- *Developing and comparing different methodological variants for the density estimation problem*

- *Exploiting Secure Multi-party Computation (SMC) for multi-MNO data fusion*
  - capacity building, technical and legal aspects


LBS
RAN-SIG
CN-SIG
CDR
LBS
controller
CDR
Billing System
Core Network (CN)
Radio Access Network (RAN)

MNO data



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

RMF - hourglass model

Statistics
S-Layer

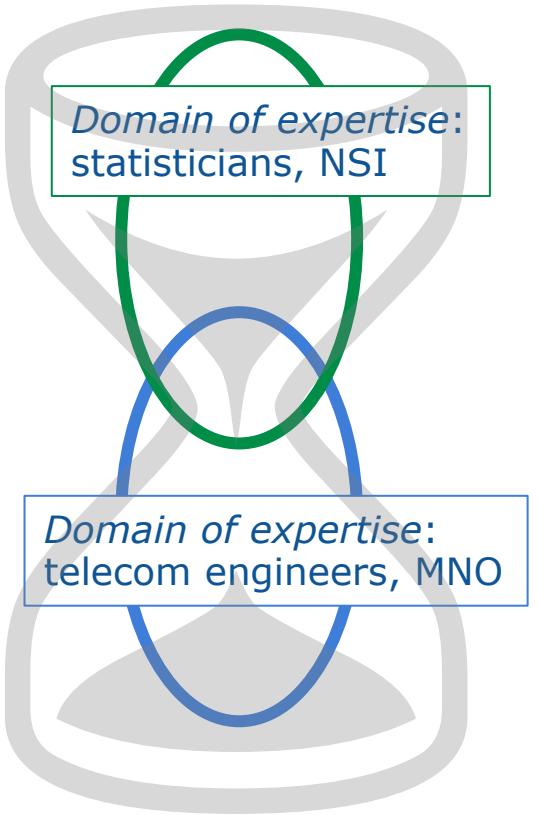
Definitions of statistical indicators
- Heterogeneity of applications & use-cases
- Diversity of statistical definitions
- Complexity of statistical objects
- Multiple NSIs

Convergence
C-layer

Few common definitions

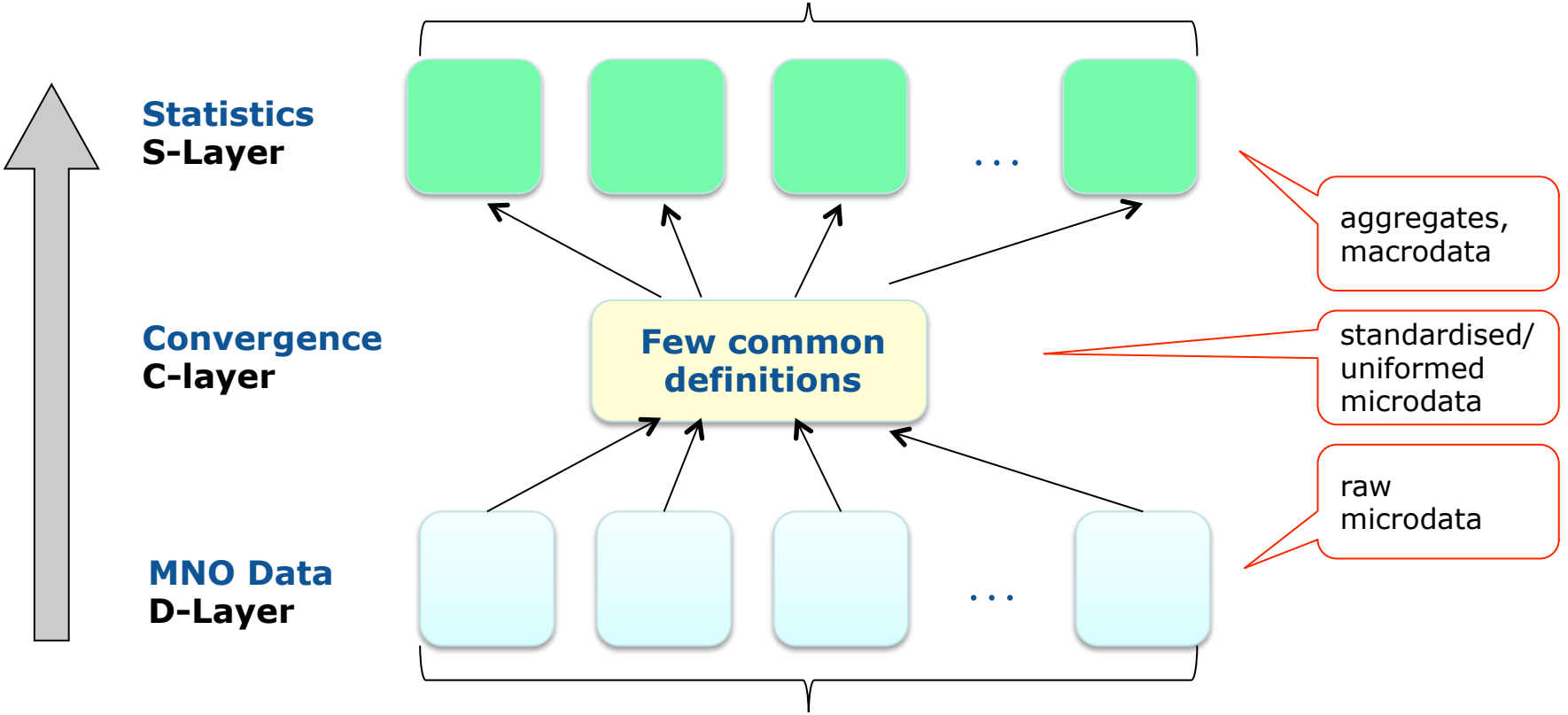
MNO Data
D-Layer

Embeds raw MNO data
- Data Heterogeneity
- Diversity of data collection methods
- Complexity of data semantics
- Multiple MNOs



RMF - hourglass model

Multiple data users: EUROSTAT, 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

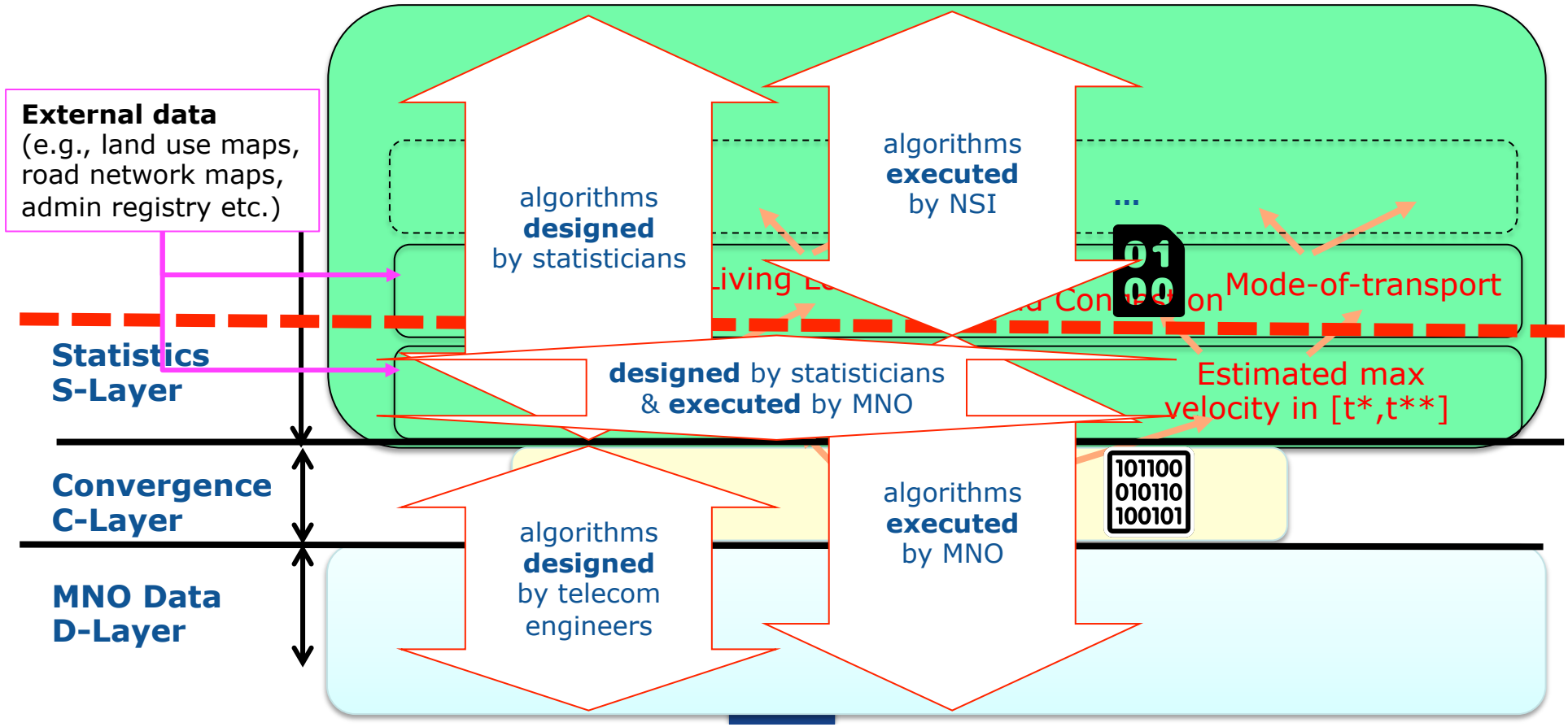
- ***Decouples the complexity & heterogeneity of the 2 domains***
  - ⇨ hides complexity & heterogeneity of MNO data to statisticians
  - ⇨ hides complexity & heterogeneity of statistical concepts to MNO engineers
- ***Decoupling allows for independent development, adoption and evolution at each domain***
- ***C-layer*** = *abstract "knowledge interface" between domains*

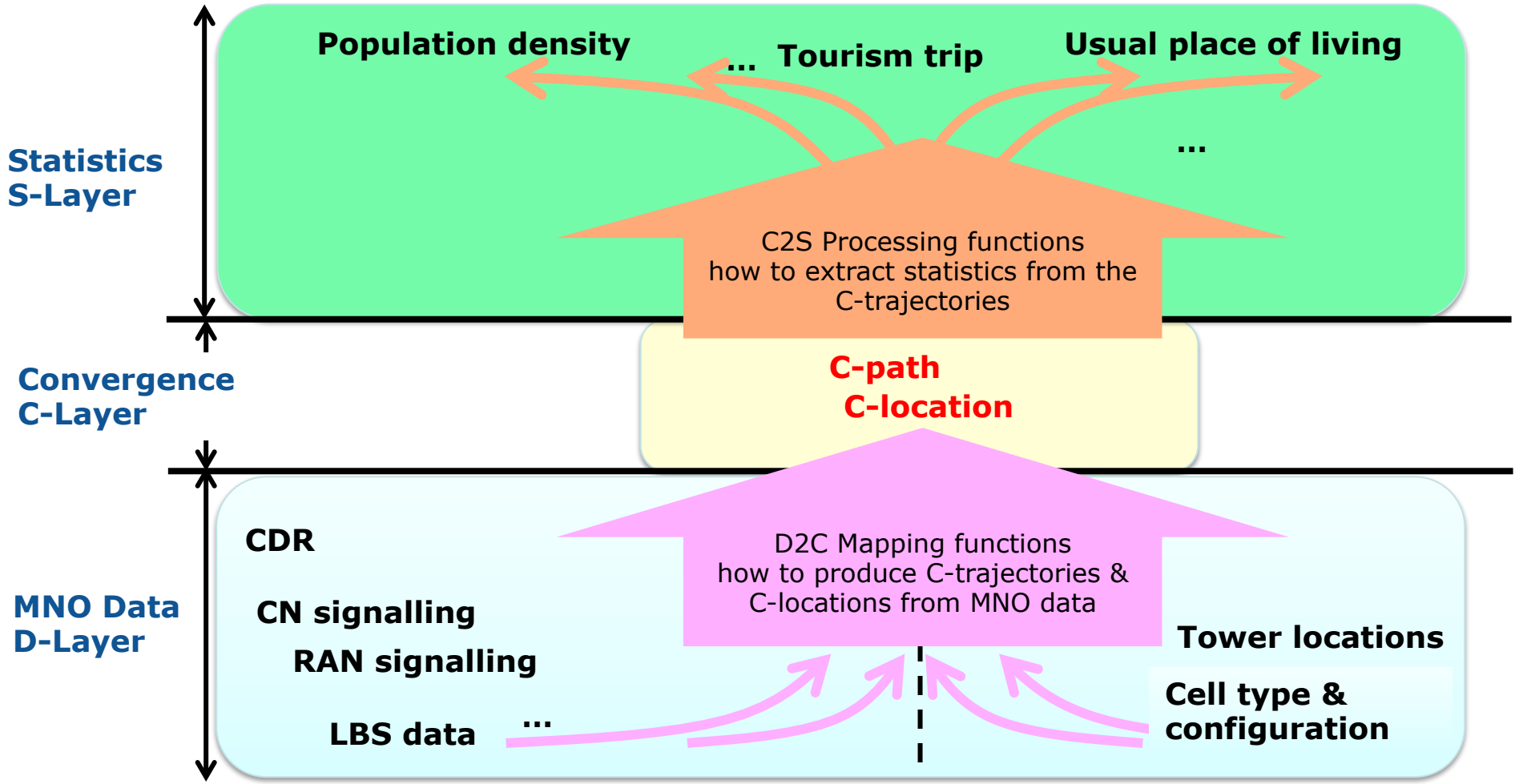
  = *"common language" for the different actors across the two domains during the algorithm design phase*

  ≠ *physical interface for data export!*

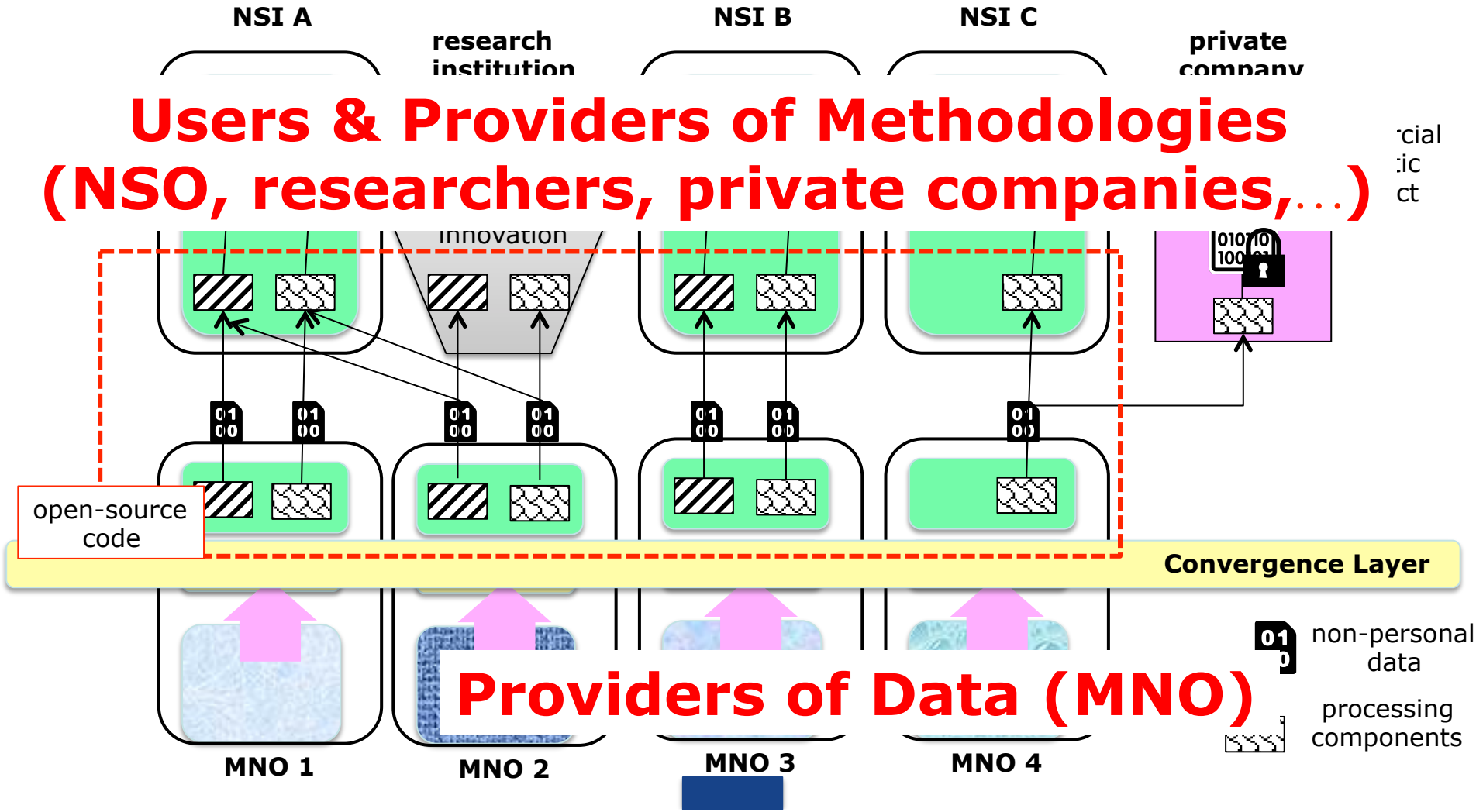
Decouple the DESIGN from the EXECUTION of algorithm:
roles and interfaces in design phase ≠ roles and interfaces in execution phase

Algorithm design versus execution




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-path
C-location
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

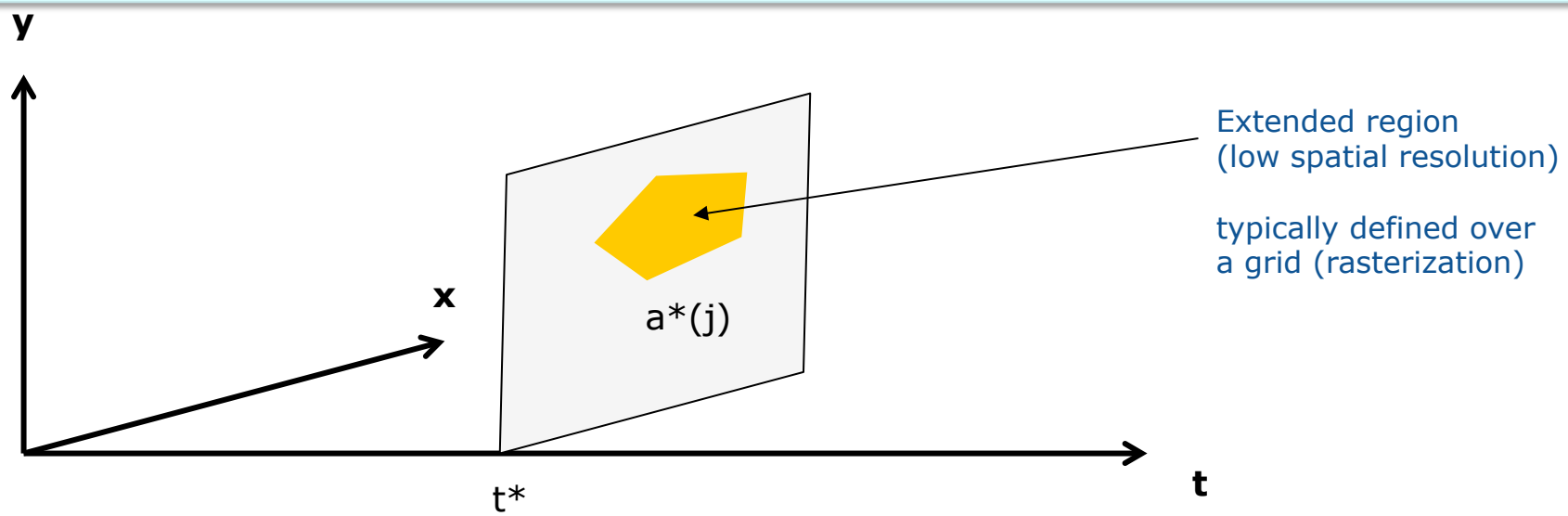
C-layer as a common substratum for MNO data users



C-location

x,y	spatial dimensions (latitude, longitude)
t*	reference time
a*(j)	best-guessed bounding area at time t* for mobile user j

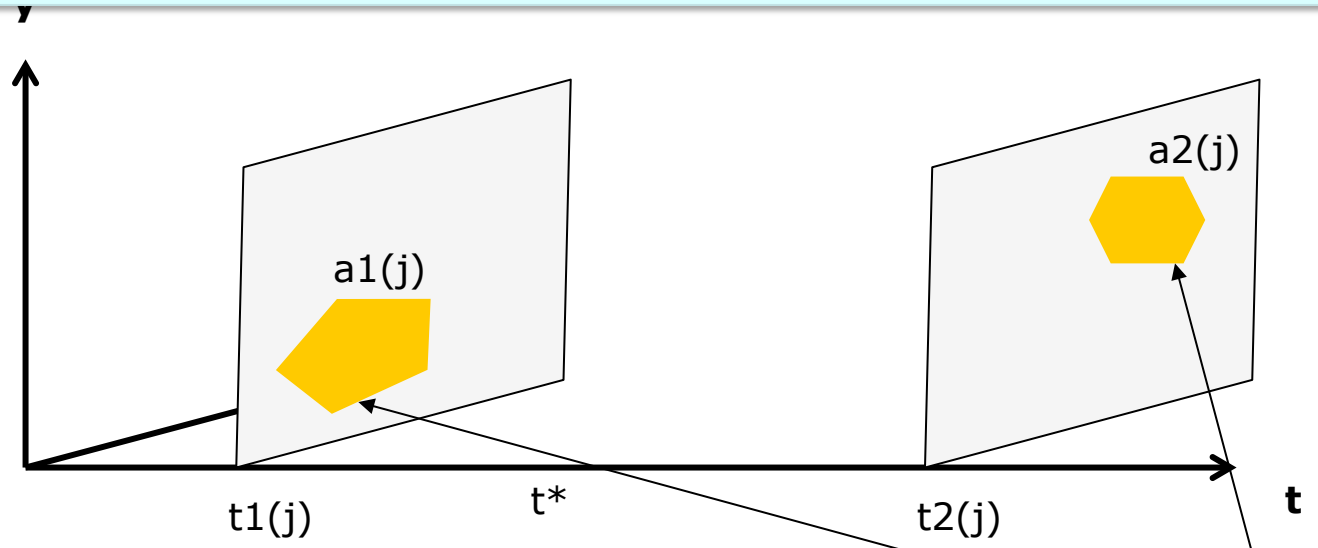
Spatial (low) resolution: you do not have a point, but an extended area that is likely to contain the actual point



C-path

x,y spatial dimensions (latitude, longitude)
t1(j), t2(j) observation instants for mobile user j
a1(j), a2(j) best-guessed bounding area at time t1,t2 for mobile user j

Temporal (low) resolution: the desired reference time t might not be included in the set of available observation times*



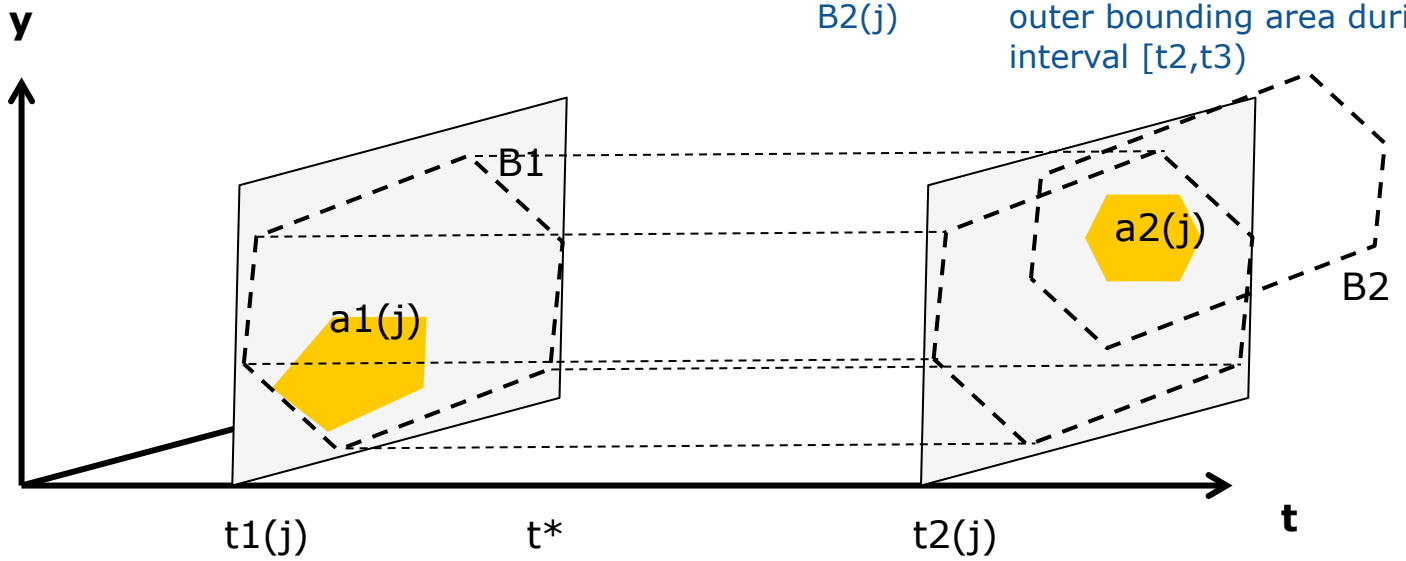
t1(j) and t2(j) are associated to individual mobile user j. They are dictated by the MNO network configuration, types of MNO data etc. logically they refer to the lower D-layer

t* is the reference time set by the S-layer expert (statistician), common for all mobile users (not depending on j)

Observed or at sparse sampling times (low temporal resolution)

C-path

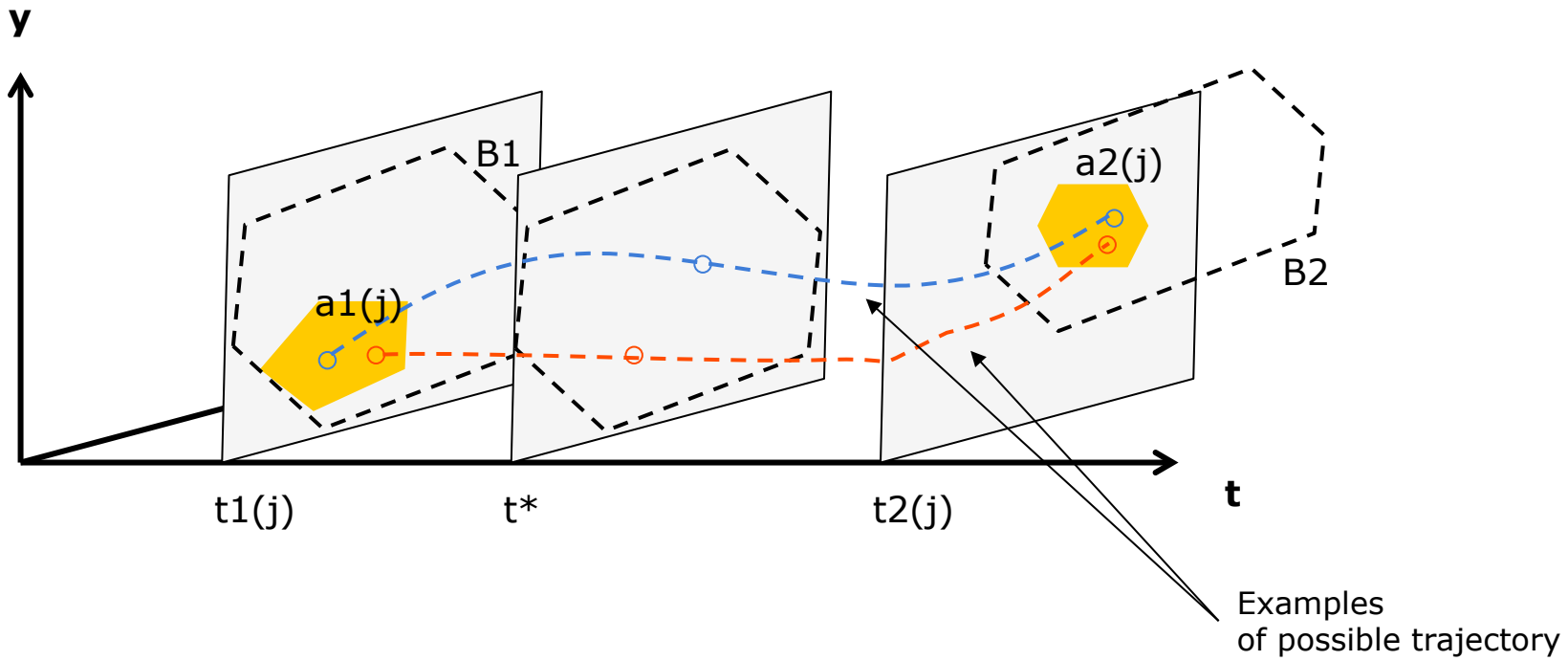
x,y	spatial dimensions (latitude, longitude)
t1(j), t2(j)	observation instants for mobile user j
a1(j), a2(j)	best-guessed bounding area at time t1,t2 for mobile user j
B1(j)	outer bounding area during the interval [t1,t2) (→Location Area in GSM)
B2(j)	outer bounding area during the interval [t2,t3)



However, if MNO signalling data are available at the D-layer, we can identify an outer region bounding the /unknown) trajectory of the mobile users between two consecutive observation times (this is given by the Location Area, Routing Area, Paging Area in 2G/3G/4G…).

C-path

x,y	spatial dimensions (latitude, longitude)
t	time
a1,a2	best-guessed bounding area at time t1,t2
B1	outer bounding area during the interval [t1,t2) (→Location Area in GSM)

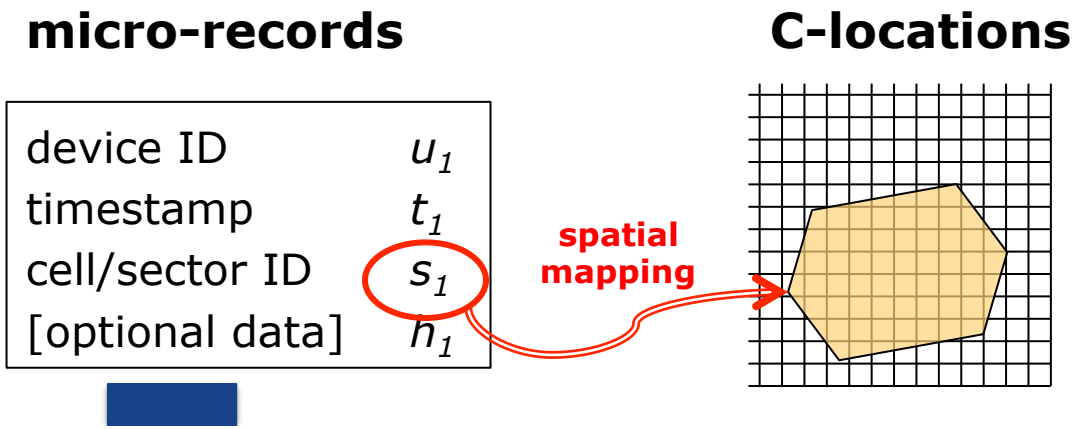


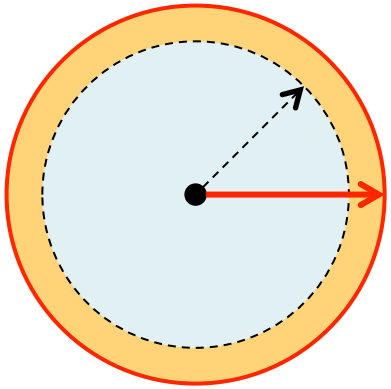
Which spatial mapping approach ?

- *Options*
  - Voronoi tessellation ☹☹
    as in most published literature
  - Better variant of Voronoi tessellation 😐
    proximus "Technology-Agnostic Cell (TAC) area"
  - Best Service Area (BSA) tessellation 😐
    ISTAT-WIND research

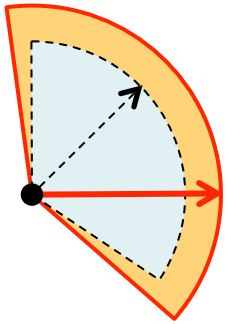
tessellations
(non overlapping)

  - Nominal coverage area (overlapping areas) ☺
    - uniform
    - non-uniform (mobloc)

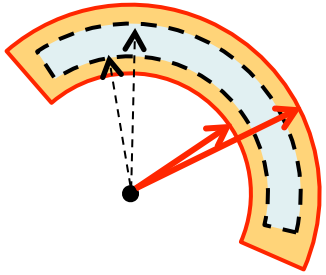




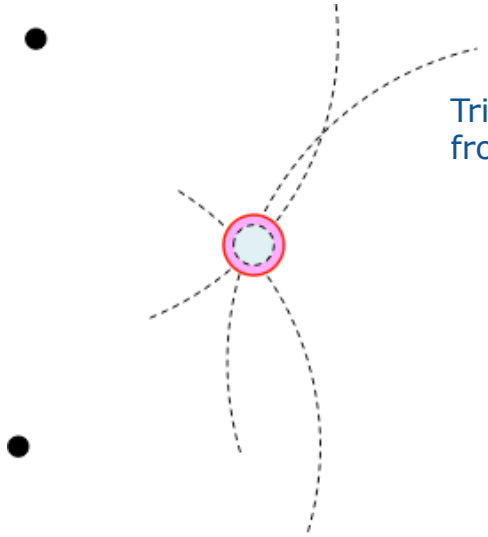
circular cell



120° sector cell



120° sector cell
+ range data
(Time Advance)



Triangulation
from LBS system

Tower position $\boldsymbol{q}_k$

Nominal cell radius

Nominal cell coverage area

C-location radius r_k

C-location a_k

Current activities on MNO data in Eurostat

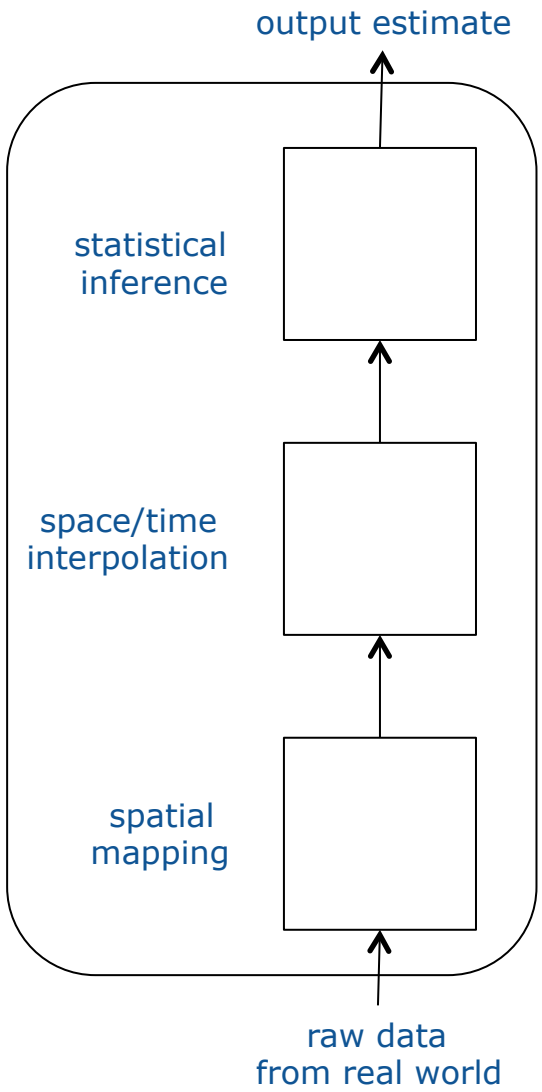
- *Developing Reference Methodological Framework (RFM) for using MNO data for Official Statistics*
  - focus on presence & mobility patterns
  - CDR and signalling data

- *Developing and comparing different methodological variants for the density estimation problem*

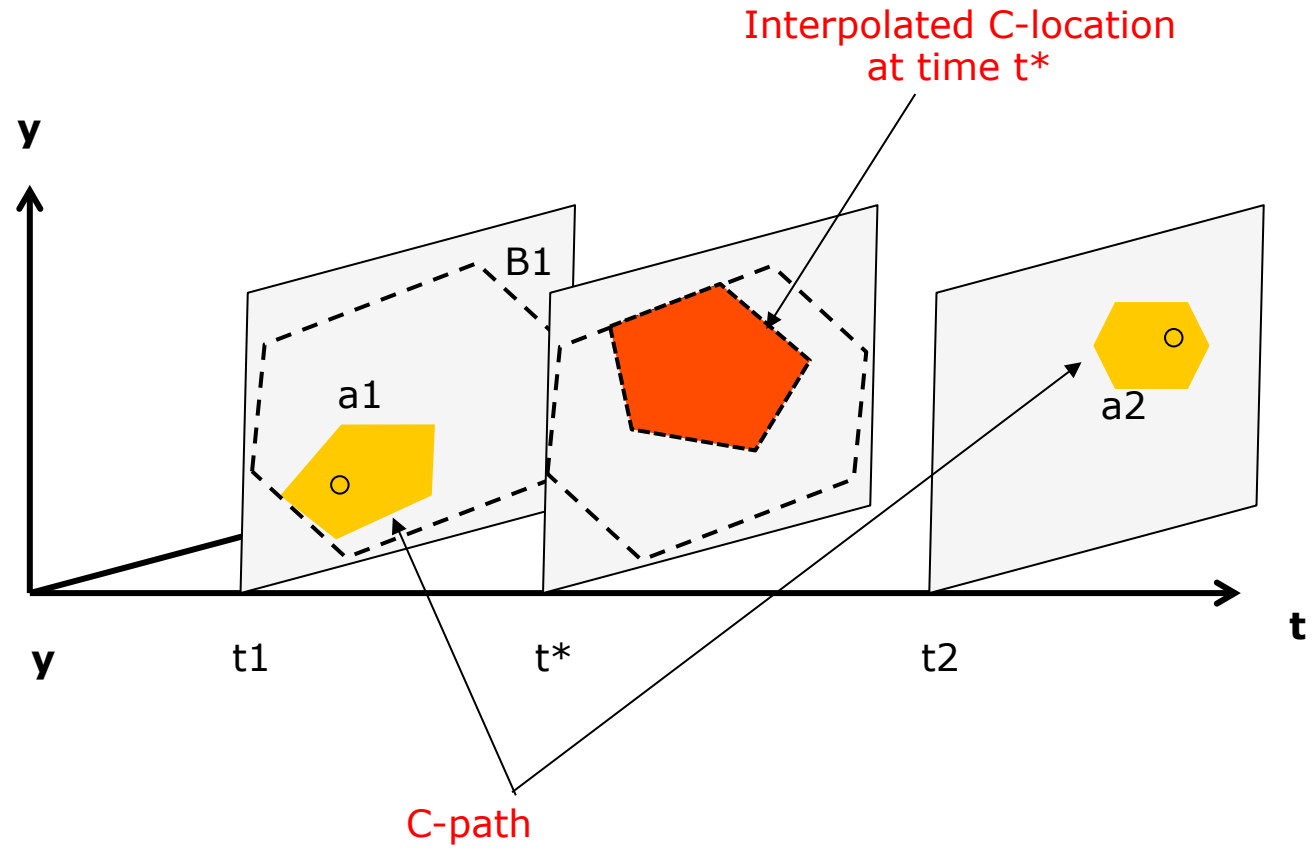
- *Exploiting Secure Multi-party Computation (SMC) for multi-MNO data fusion*
  - capacity building, technical and legal aspects

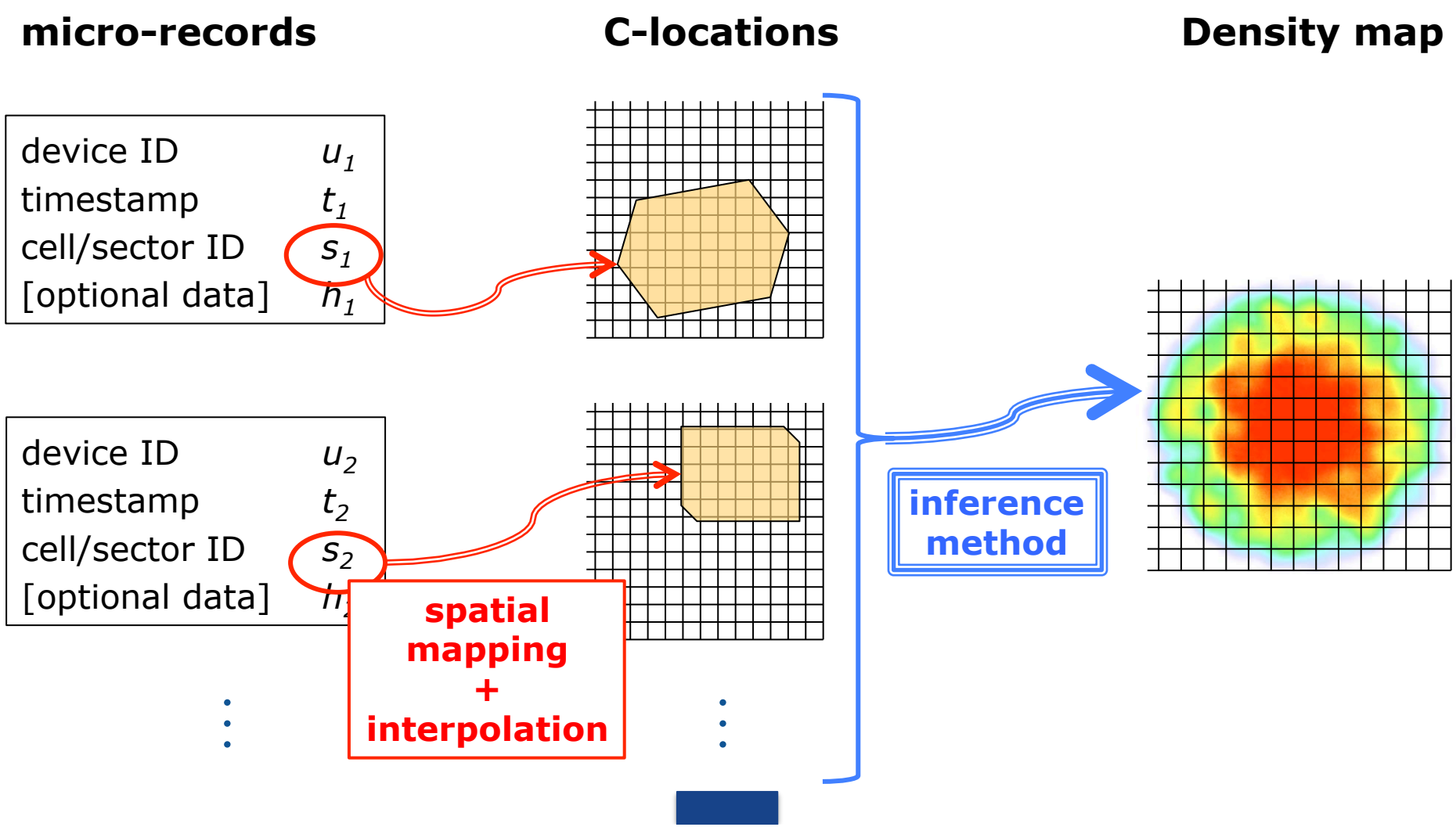
Design choices

- *1) Which spatial mapping approach ?*
  - Voronoi tessellation ☹☹
  - Variant of Voronoi tessellation 😐
  - Best Service Area (BSA) tessellation 😐
  - Overlapping coverage area ☺☺
    uniform or non-uniform ?
- *2) Which space/time interpolation method?*
  - Zero-order interpolation
  - …
- *3) Which inference method?*
  - Area Proportional
  - Maximum Likelihood
  - Hierarchical Bayesian
  - …



Spatio/Temporal Interpolation




micro-records
C-locations
Density map
device ID u_1
timestamp t_1
cell/sector ID s_1
[optional data] h_1
device ID u_2
timestamp t_2
cell/sector ID s_2
[optional data]
spatial mapping + interpolation
inference method

Current activities on MNO data in Eurostat

- *Developing Reference Methodological Framework (RFM) for using MNO data for Official Statistics*
  - focus on presence & mobility patterns
  - CDR and signalling data

- *Developing and comparing different methodological variants for the density estimation problem*

- *Exploiting Secure Multi-party Computation (SMC) for multi-MNO data fusion*
  - capacity building, technical and legal aspects

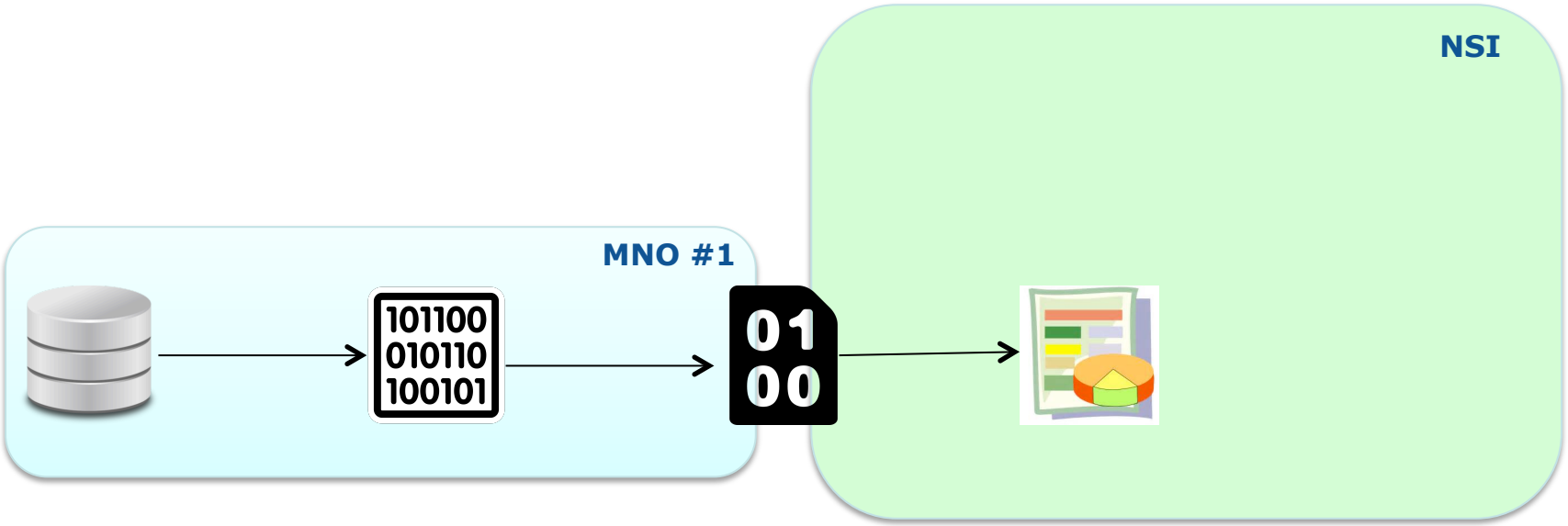
Stage 1 scenario

Raw micro-data (D-layer)

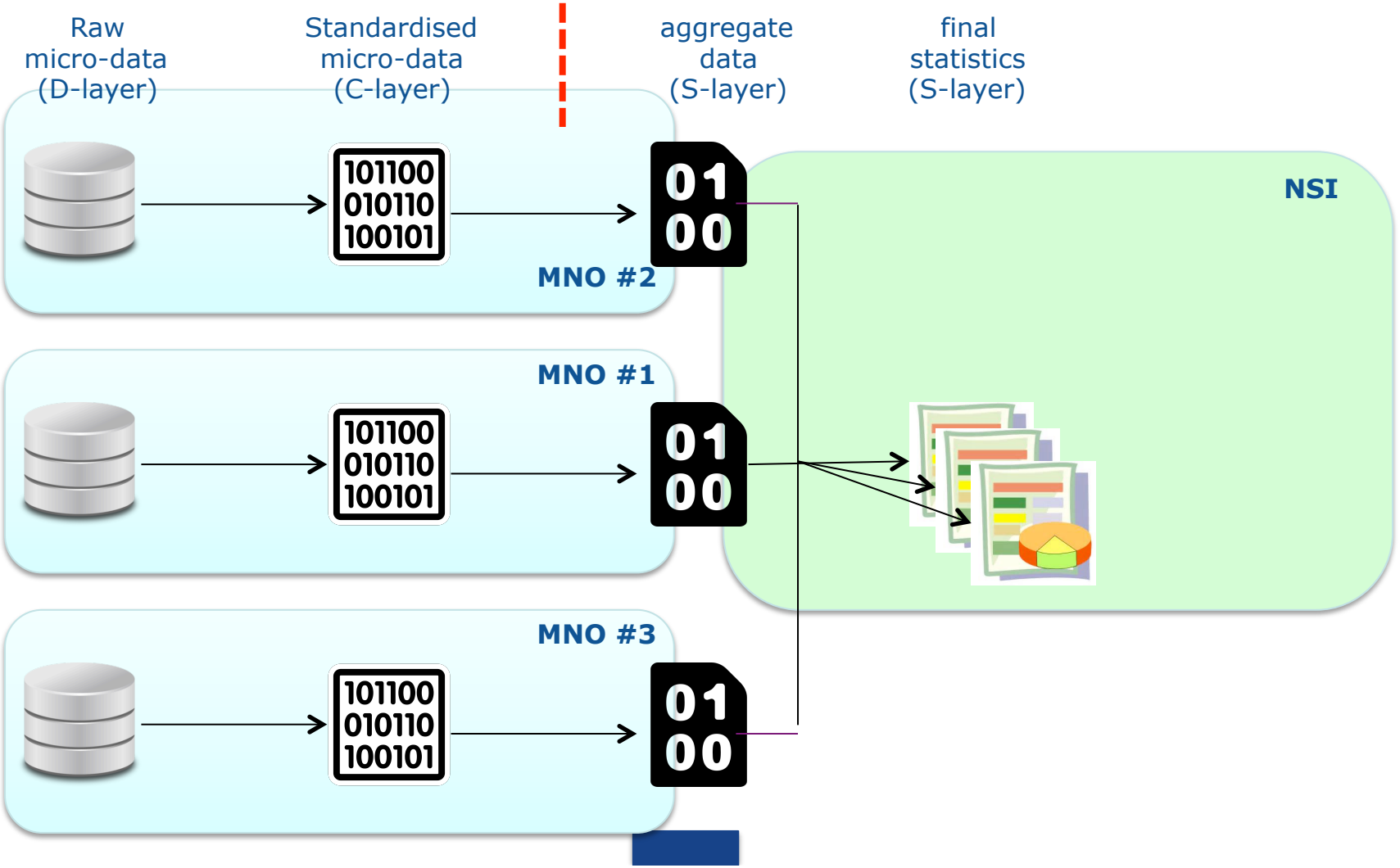
Standardised micro-data (C-layer)

aggregate data (S-layer)

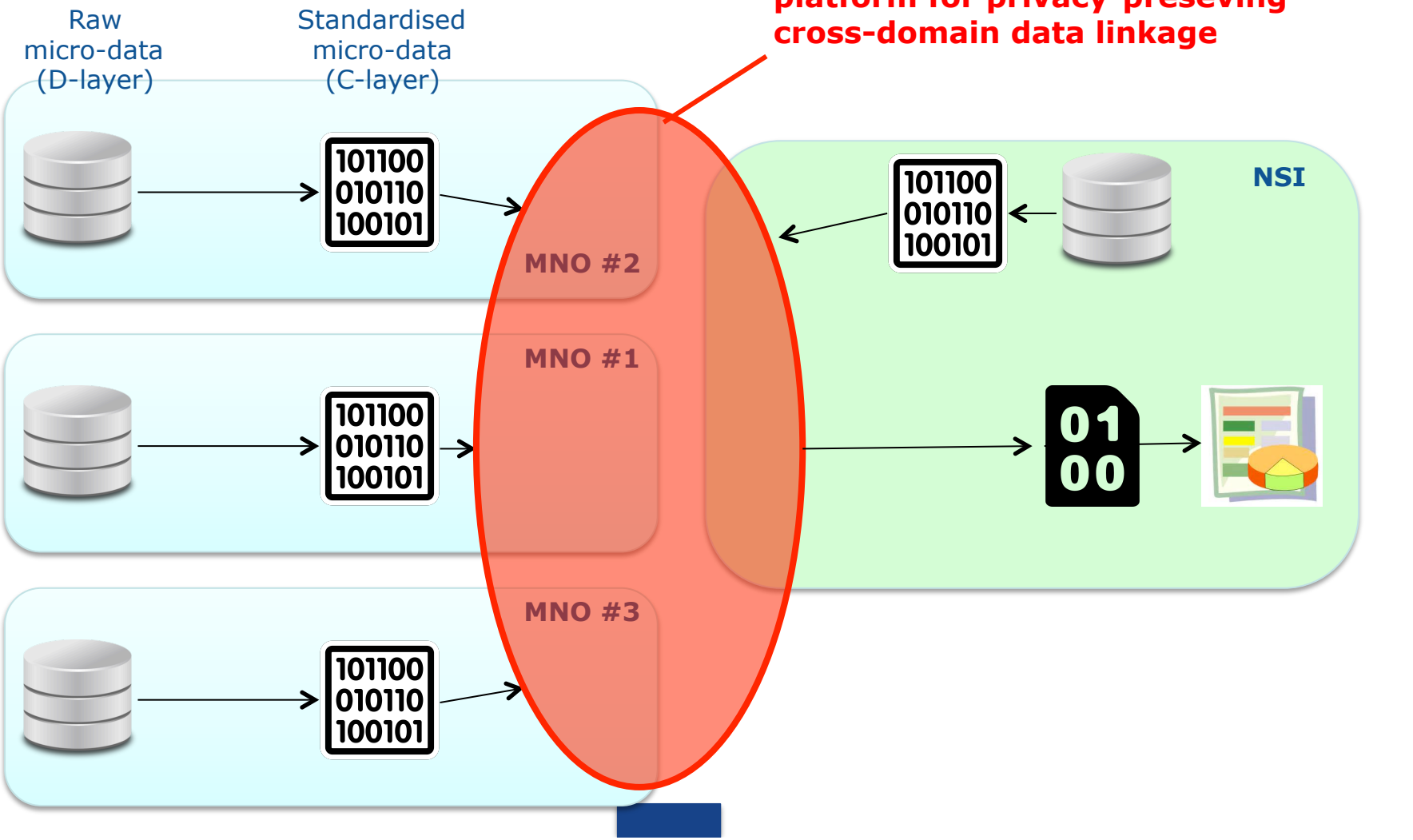
final statistics (S-layer)



Stage 2 scenario

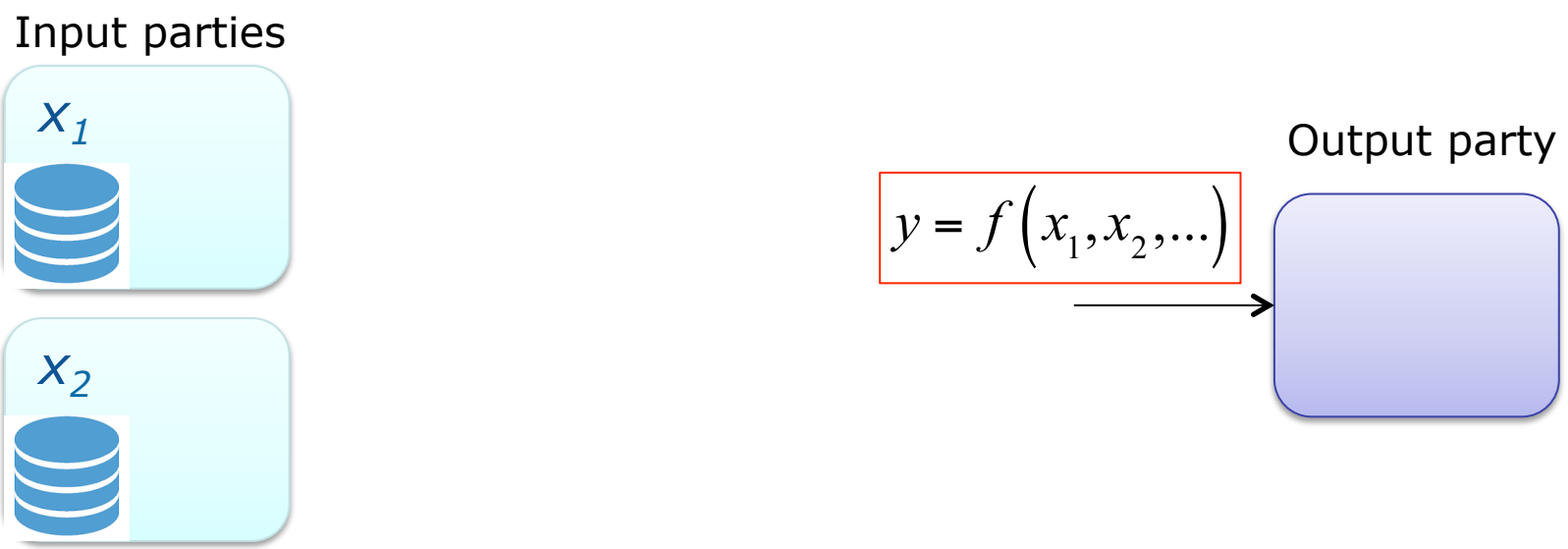


Stage 3 scenario



Problem statement

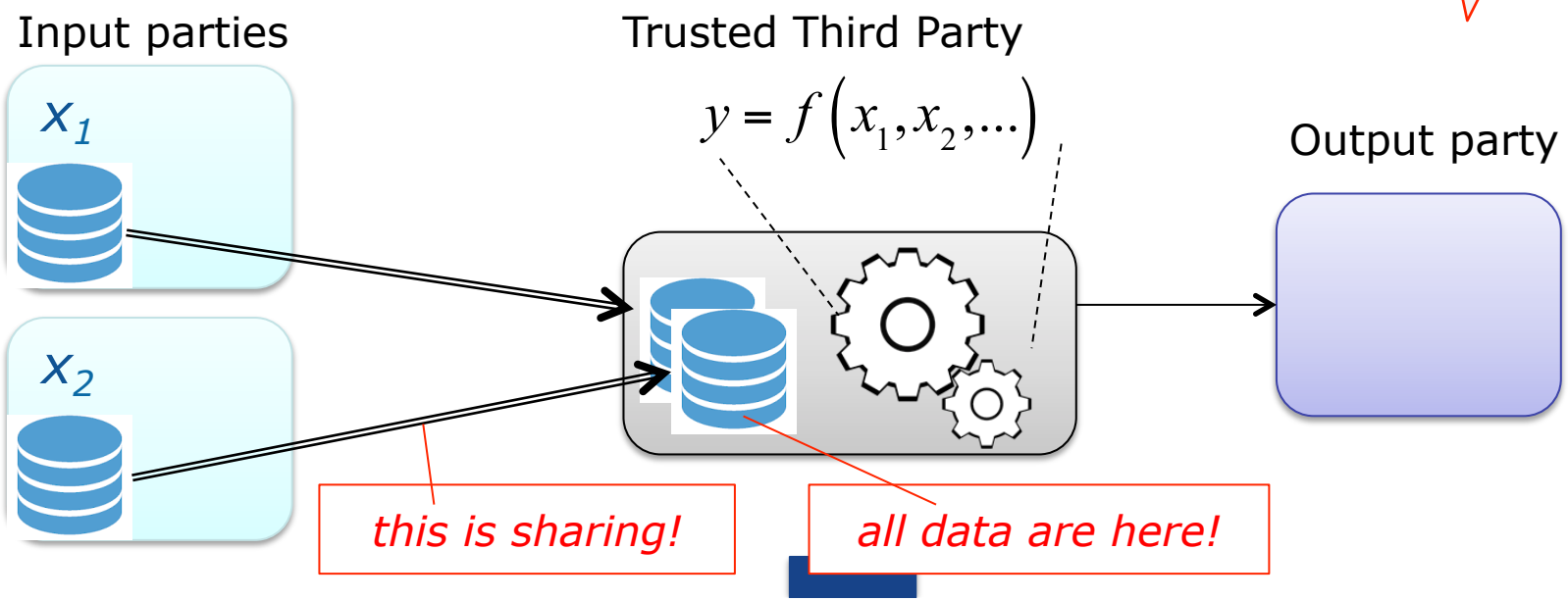
- Two or more **input parties** held confidential data x_1, x_2, …
- They agree that another **output party** (e.g., the Stat. Office) computes the result of a statistical function $y=f(x_1,x_2,…)$ on their input data (e.g., regression coefficients)
- Each input party does not want to disclose its input data with the other input/output parties



Trusted Third Party (TTP)

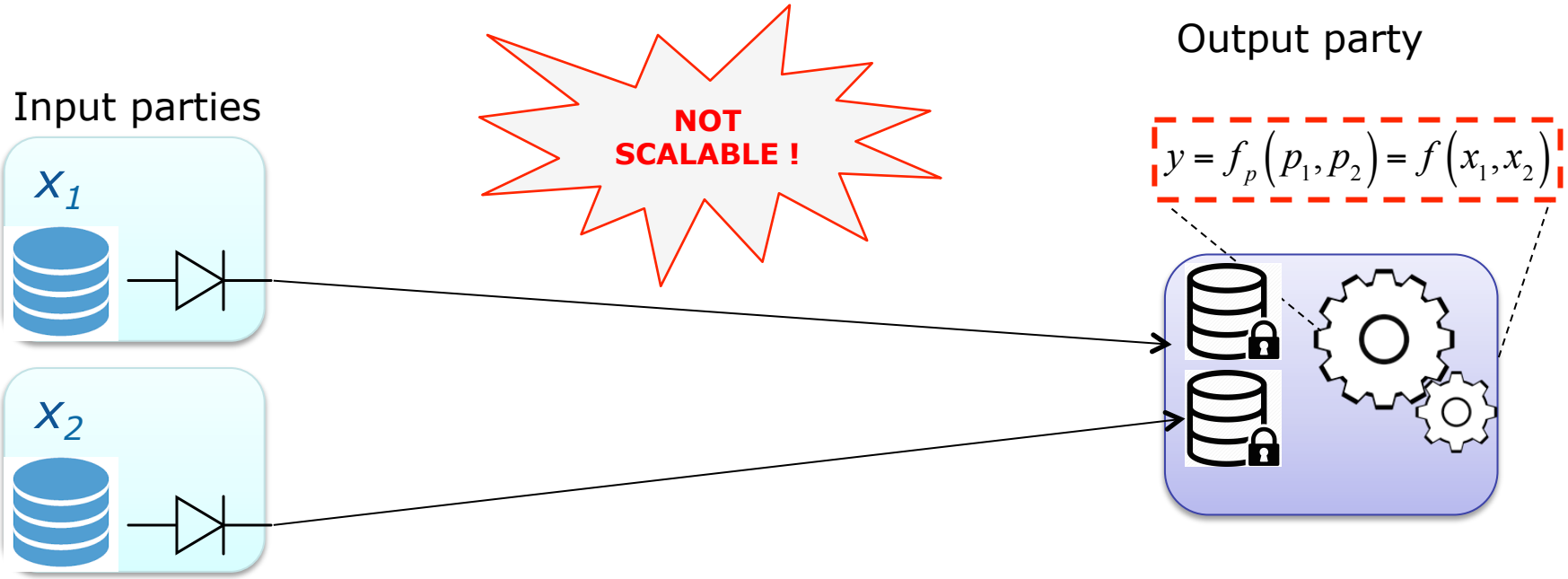
- data sharing still occurs (with TTP)
- risk concentration at TTP
- a single entity trusted by all input/output parties might not exist

NOT A SOLUTION!



Fully Homomorphic Encryption (FHE)

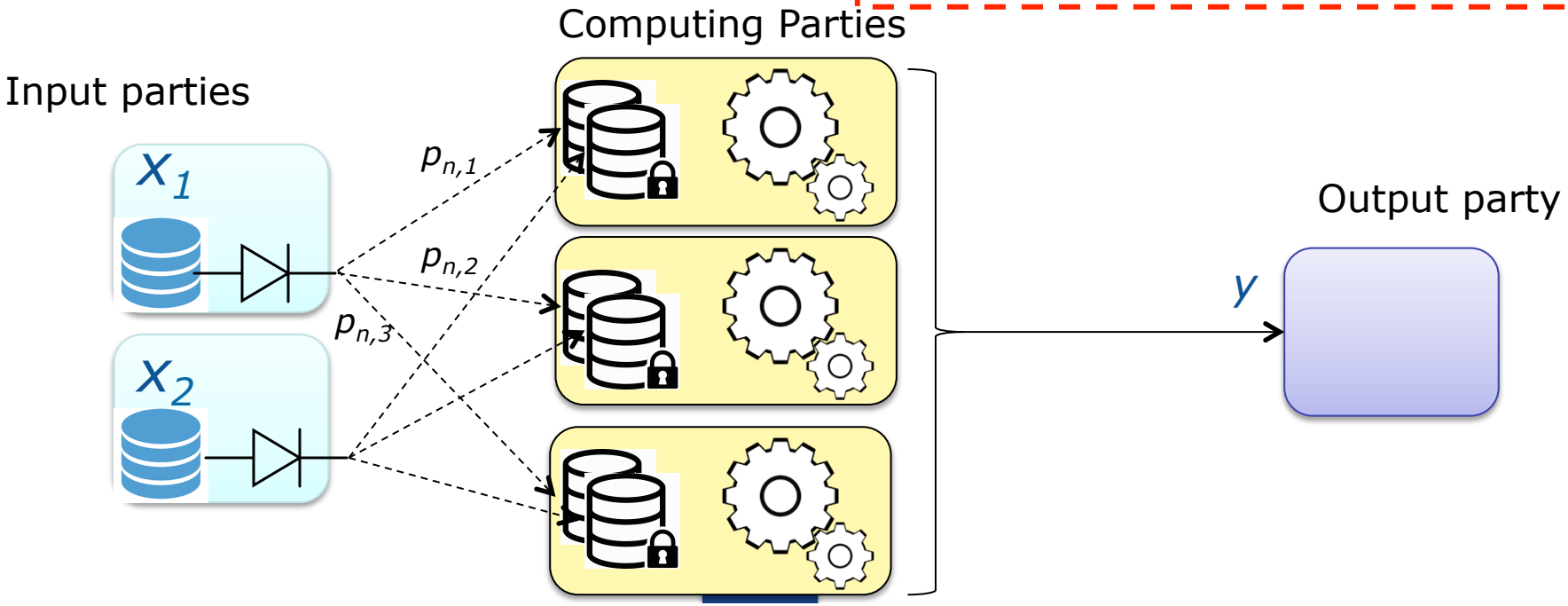
- Encrypt the input data x_n into p_n (non-reversible transformation)
- The computation on encrypted data eturns the same output value that would be obtained on the input data (homomorfism)
- → along the process the original inputs are never reconstructed (no "decryption" takes place) - only the output is revealed
- *theoretically possible, but practically unfeasible in most cases*



Secure Multi-party Computation via Secret Sharing

- *Each element of secret input x_n is transformed into K "shares" $p_{n,1}$, $p_{n,2}$ … $p_{n,k}$ that are distributed to different computing parties*
- *The computation is distributed (shared) among the computing parties*
- *The computation on secret shares returns the same output value that would be obtained from the input data (homomorfism)*

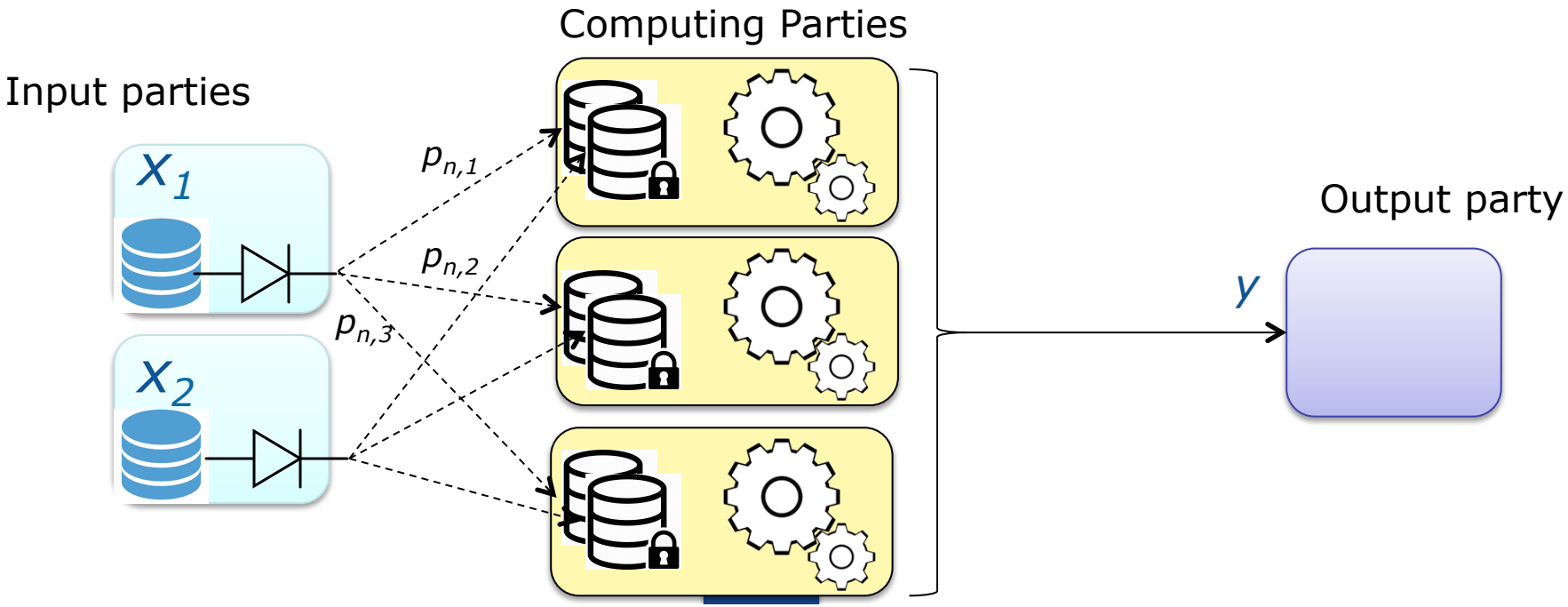
$$y = f_s\left(\left\langle p_{1,1}, p_{1,2}, p_{1,3}\right\rangle, \left\langle p_{2,1}, p_{2,2}, p_{2,3}\right\rangle\right) = f\left(x_1, x_2\right)$$



Secure Multi-party Computation via Secret Sharing

- *Individual shares reveal nothing about the secret input*
  - → no single party holds "data"
    → "passing shares" ≠ "sharing data"

- *Computing parties need to be trusted **collectively, not individually***

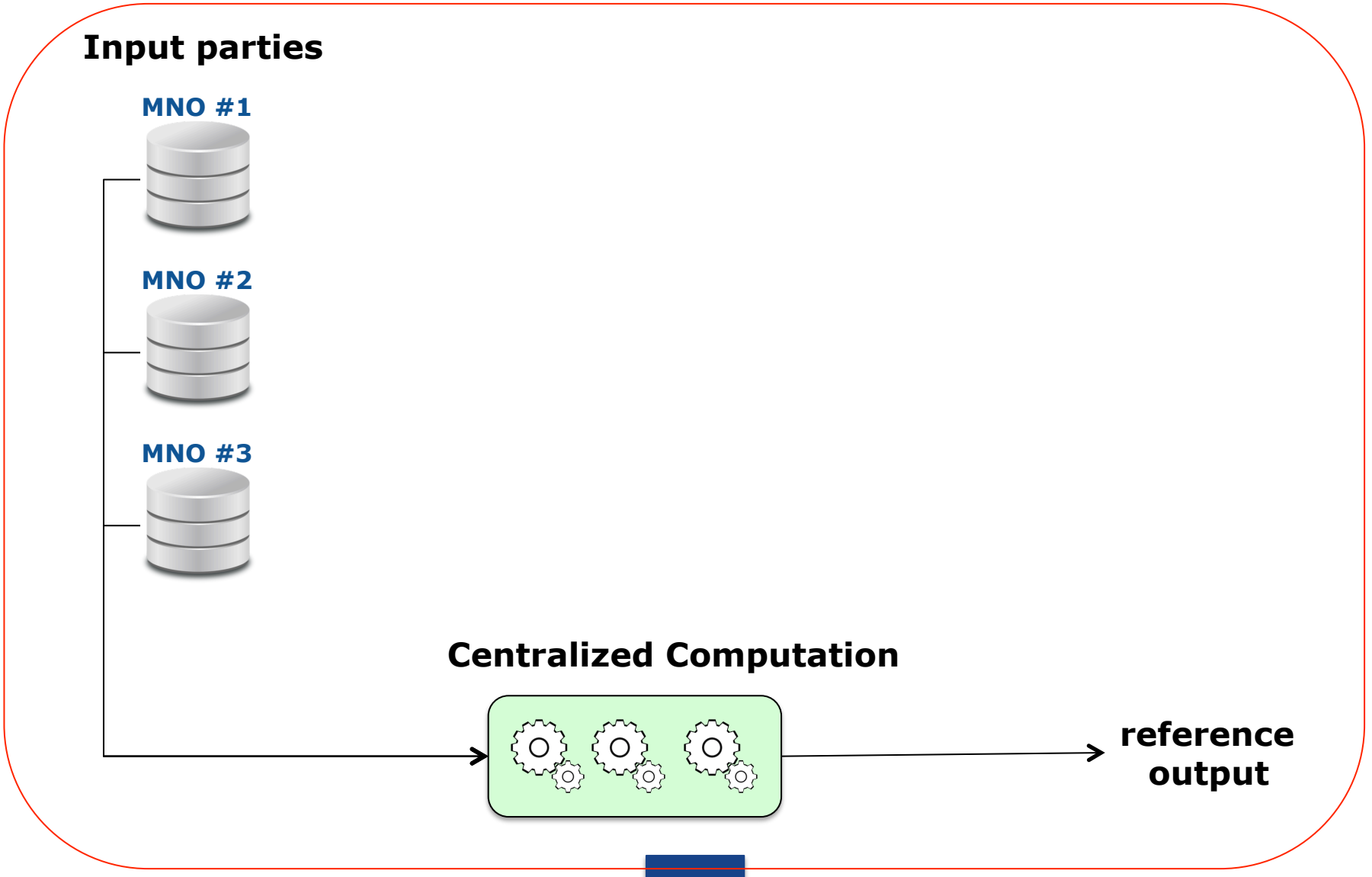


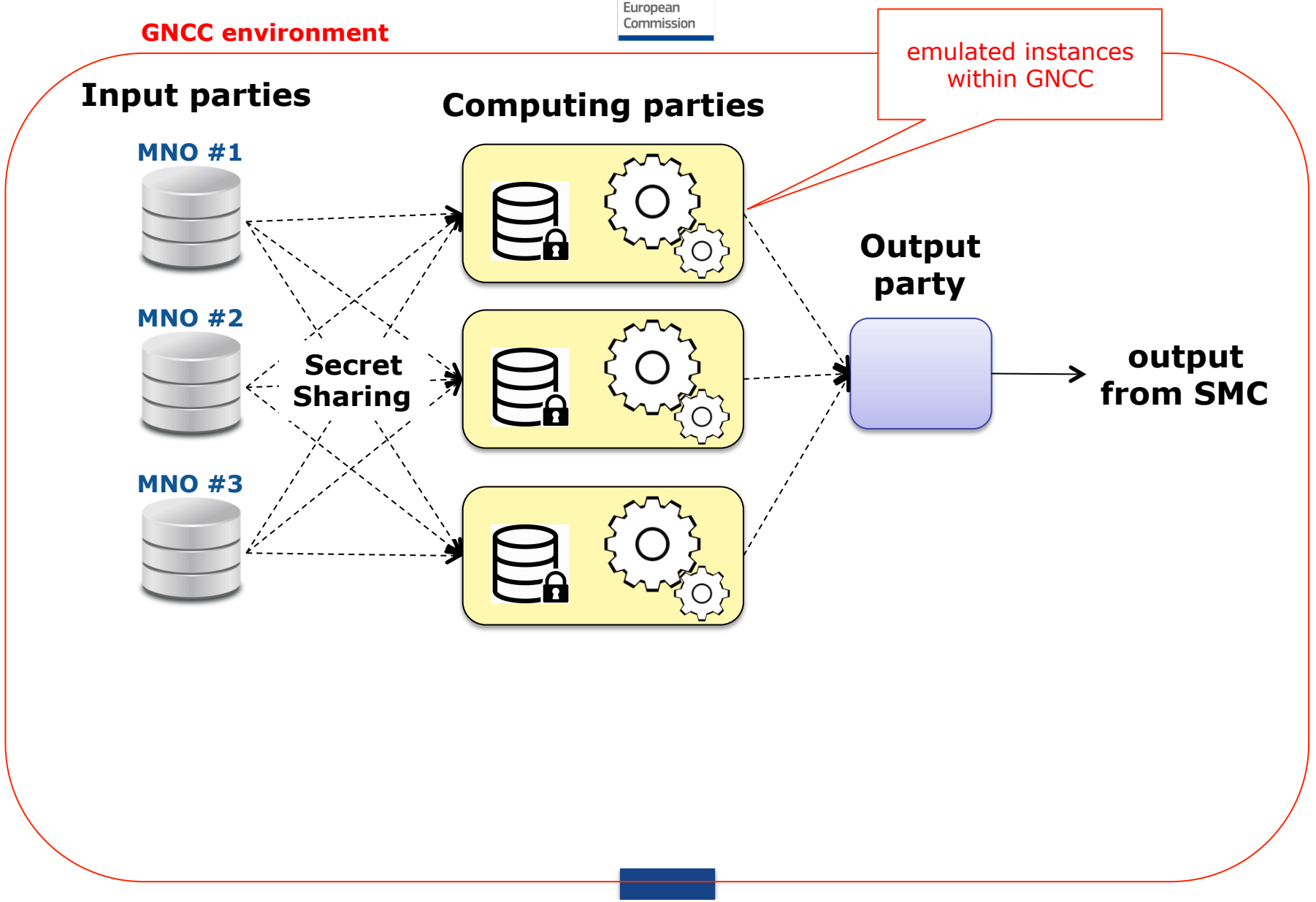
Main points of interest for this project

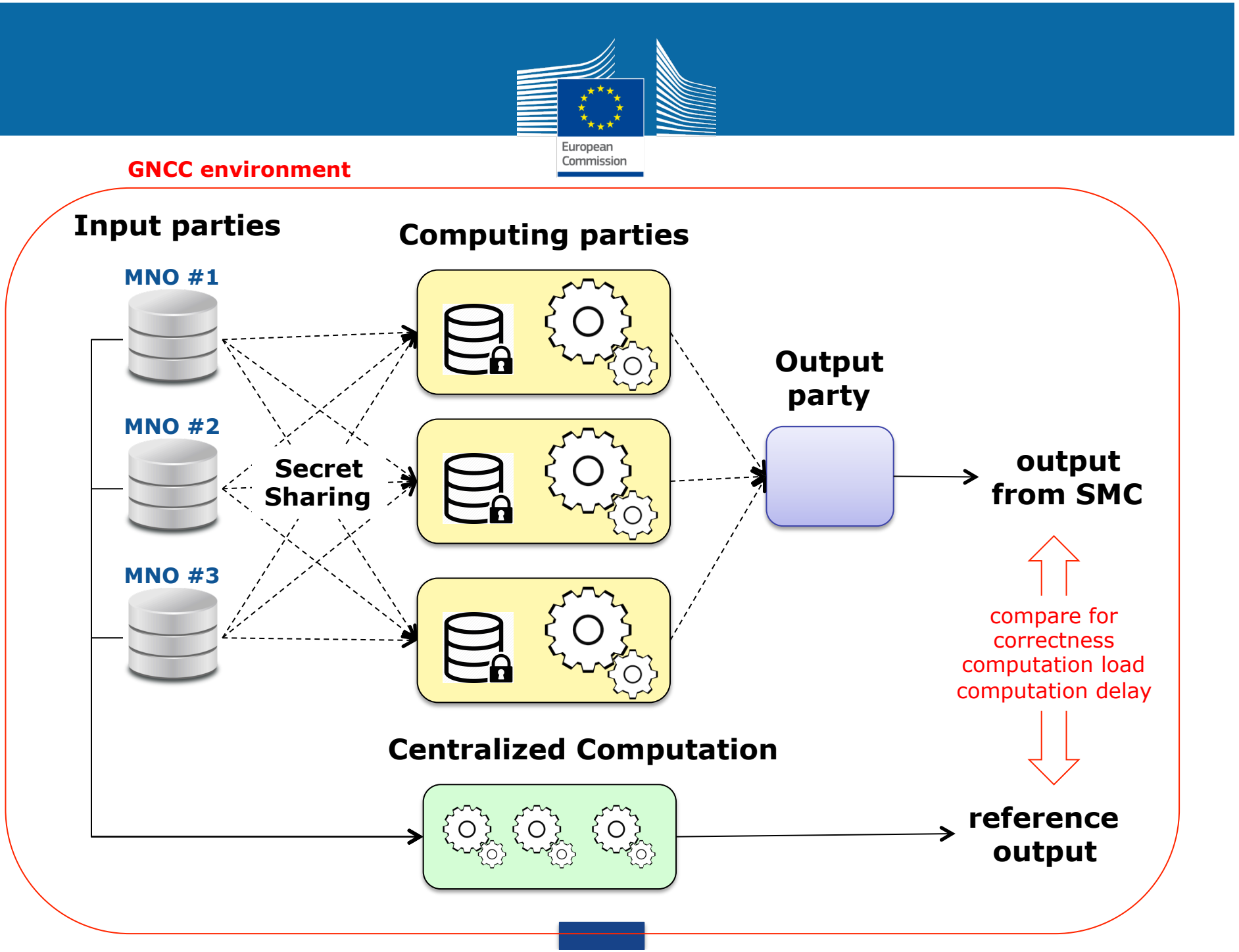
- *Develop novel **methodologies**, explore use-cases, with special focus on **multi-MNO aspects***

- *[optional] benchmark algorithm implementation in centralized vs. SMC settings*
  - verify correctness of output
  - assess computation load and delay of SMC implementation

GNCC environment




GNCC environment
emulated instances within GNCC
Input parties
Computing parties
MNO #1
MNO #2
MNO #3
Secret Sharing
Output party
output from SMC


GNCC environment
Input parties
Computing parties
MNO #1
MNO #2
MNO #3
Secret Sharing
Output party
output from SMC
compare for correctness computation load computation delay
Centralized Computation
reference output

Thanks for your attention

For follow-up:

fabio.ricciato@ec.europa.eu