



Towards a reference architecture for Trusted Smart Surveys

Fabio Ricciato

Unit B1 'Methodology; Innovation in Official Statistics'
Eurostat

Fabio.Ricciato@ec.europa.eu

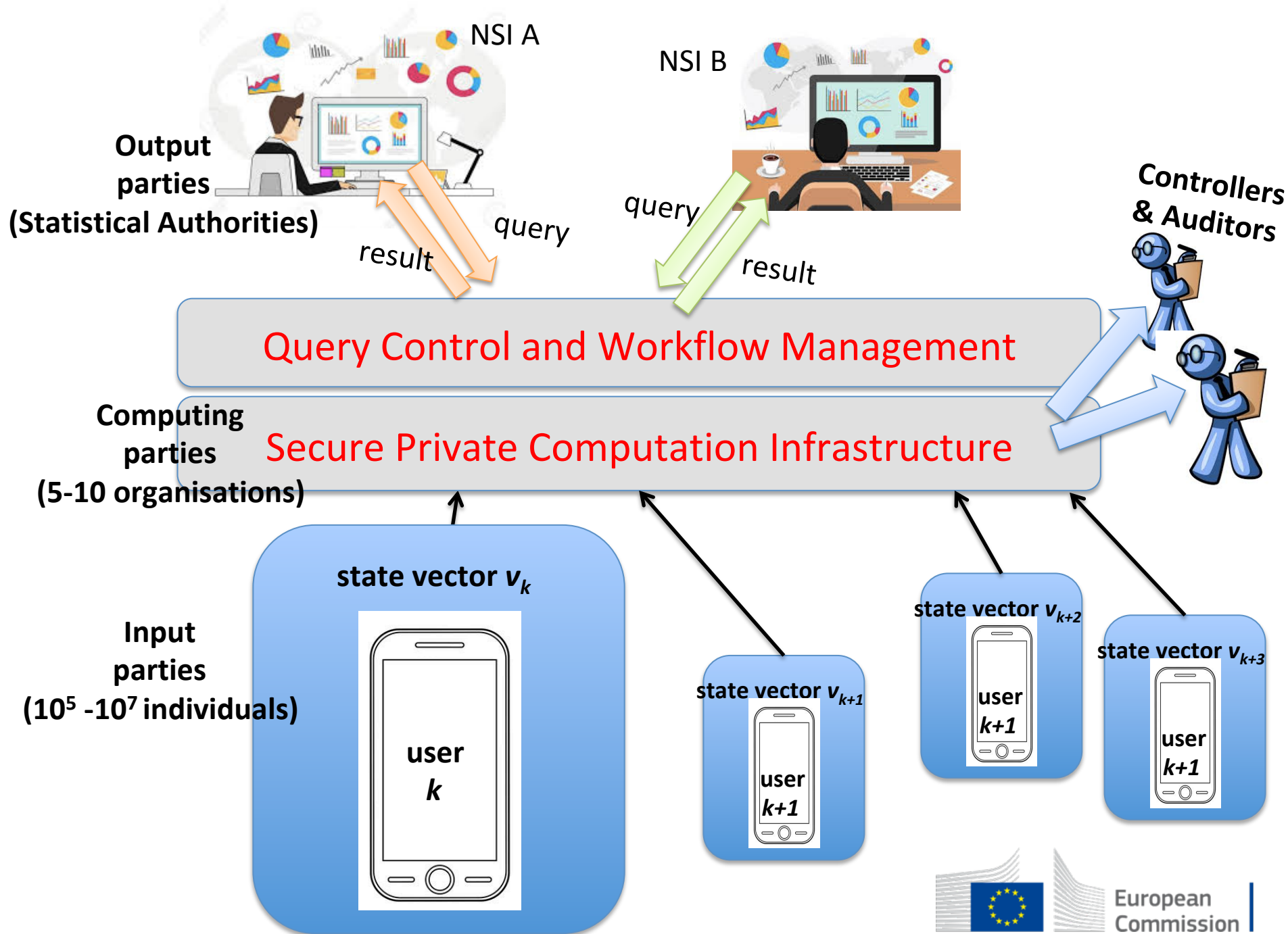
Meeting of UNECE Input Privacy Project
22 January 2021

Traditional Surveys → Smart Surveys

- ***Traditional Surveys:*** learn user behaviour by asking a batch of questions to the respondent
 - **interaction model: high attention for short time**
 - **on paper, via telephone, via web ...**
 - **NSI acquire individual data for further processing**
- ***Smart Survey:*** learn user behaviour by *interacting continuously* with the respondent via her mobile device and *by pulling information from the sensor data* collected by/through the device
 - **interaction model: low attention for long time**
 - **via personal mobile device(s)**
 - **NSI acquire individual data for further processing**

Smart Surveys → Trusted Smart Surveys

- **Trusted Smart Survey:** *learn user behaviour by interacting continuously with the respondent via her mobile device and by pulling information from the sensor data collected by/through the device*
 - **interaction model: low attention for long time**
 - **via personal mobile device(s)**
 - ~~NSI acquire individual data for further processing~~
 - **Individual data never leave the private space of the individual respondent. NSI acquire only the final aggregate results of statistical processing, and only for queries that underwent an approval process.**



Result are aggregate, non personal data
(e.g. % of total users, histogram)

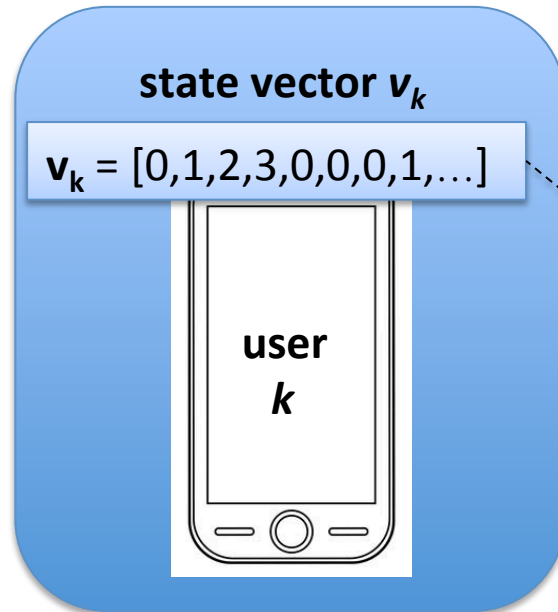
This infrastructure allows to compute the
result of pre-defined and pre-approved
function (e.g. total count, ratio, regression
coefficient, histogram, etc.) without
exposing the private input data



Computing
parties

Secure Private Computation Infrastructure

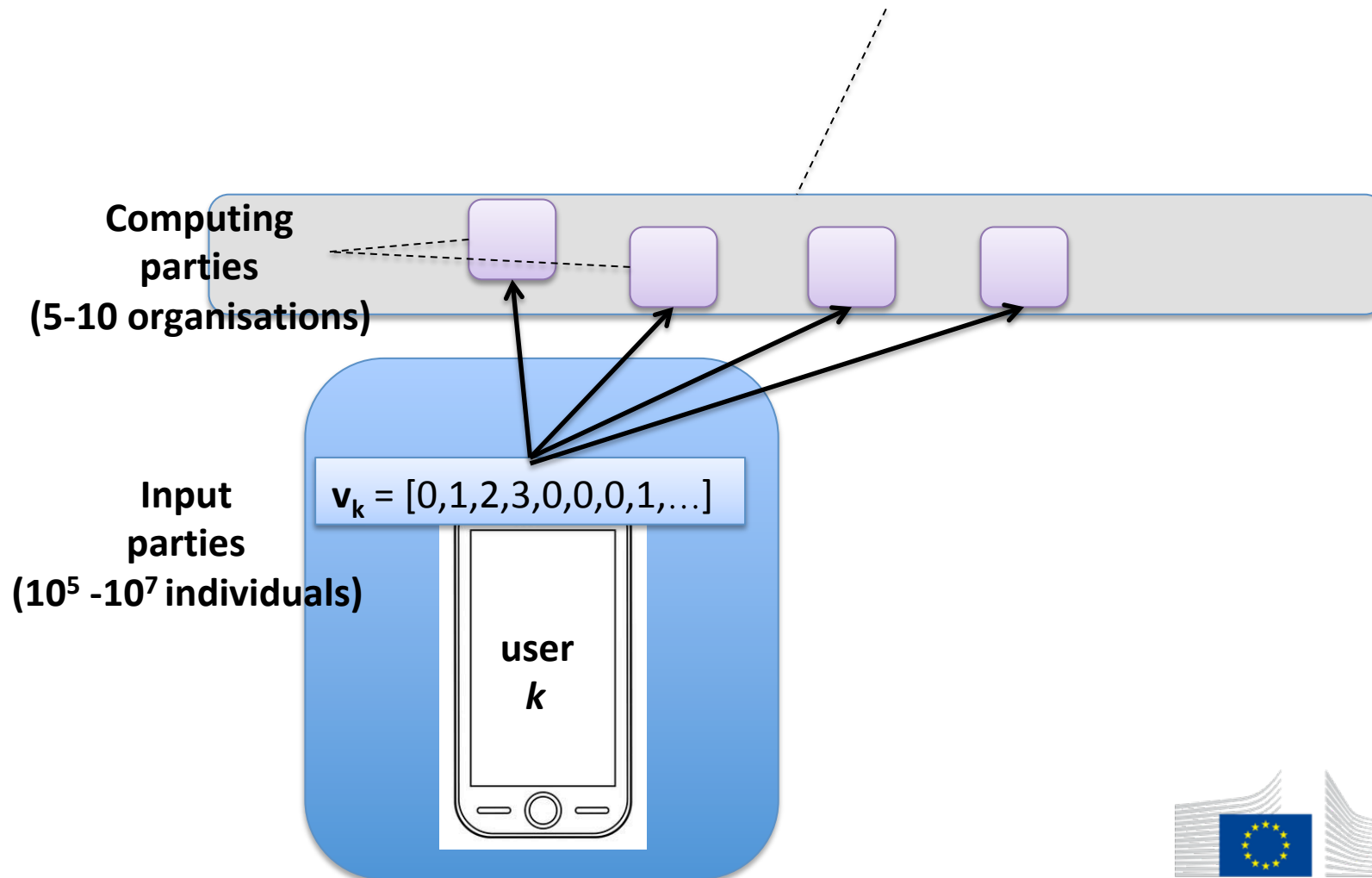
(5-10 organisations)



State Vector (SV) is a record of structured
data representing the individual user
variables.

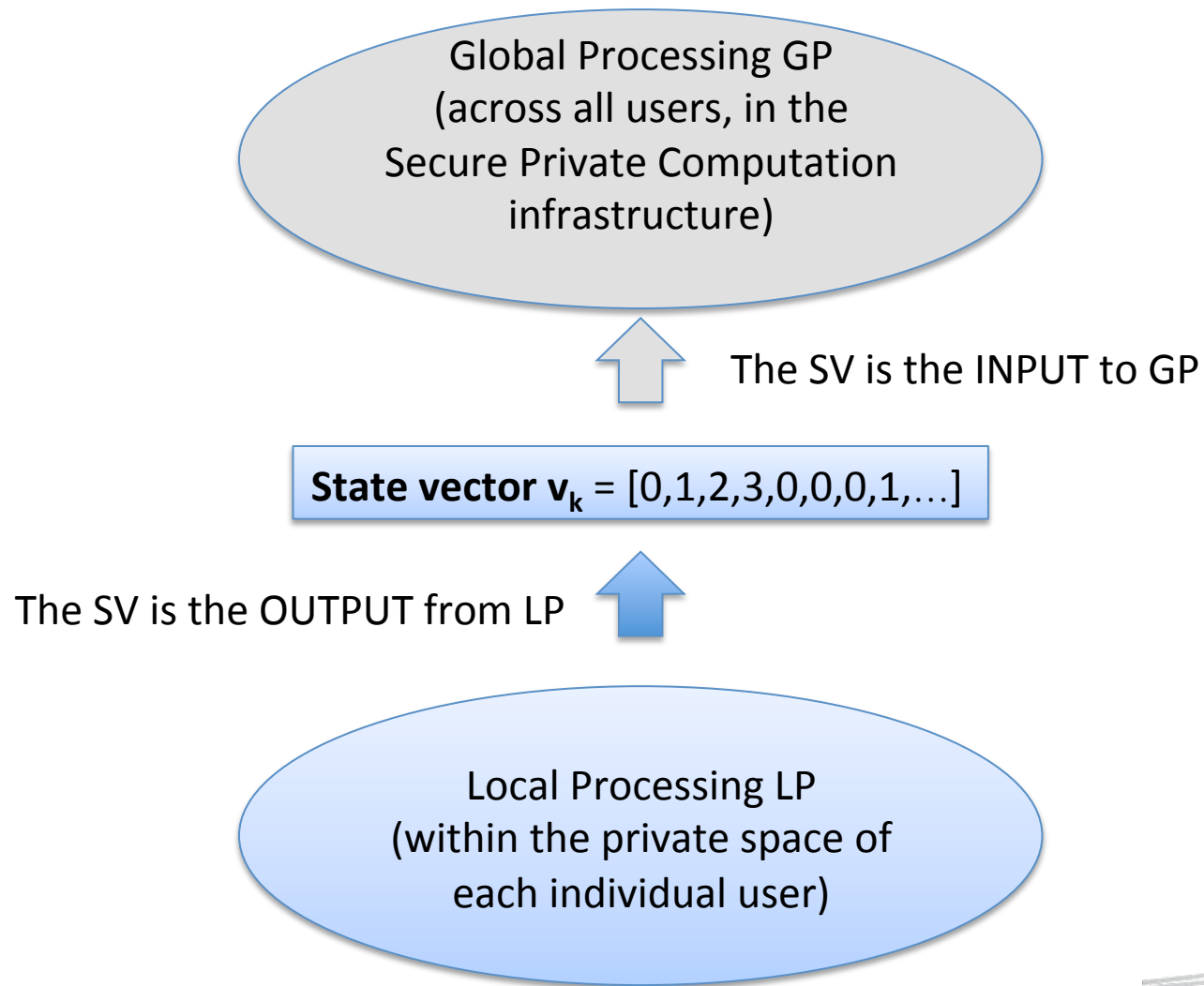
This is a piece of private input data: it never
leaves the private user space (to be defined
later).

This infrastructure could be based (also) on **Secure Multiparty Computation (SMPC)**: each SV will be “secret shared” among the computing parties. It could be based on **Trusted Execution Environment (TEE)**. It could be based on some combination of TEE and SMPC (e.g. secret sharing of TEE encryption keys,...). It could be also based on Homomorphic Encryption (HE) (although the need to use a single key pair for all individual input parties may not fit well with the requirements...)



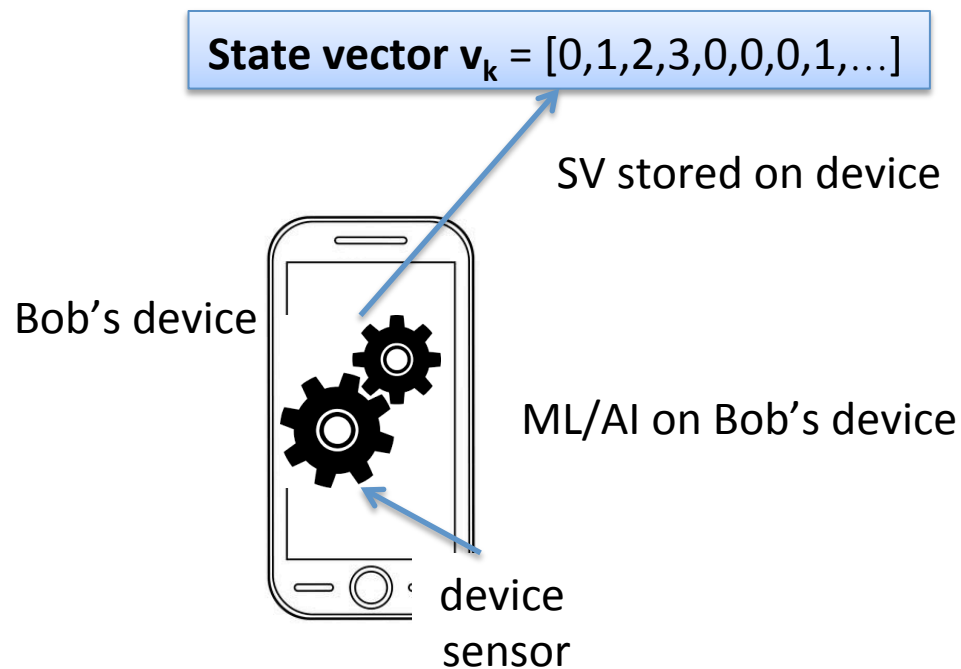
It's important to decouple the two levels of Local Processing (LP) and Global Processing (GP).

NB: You may use a "cloud" solution for GP and another cloud solution for LP, but they will be different since they need to fulfil different requirements.





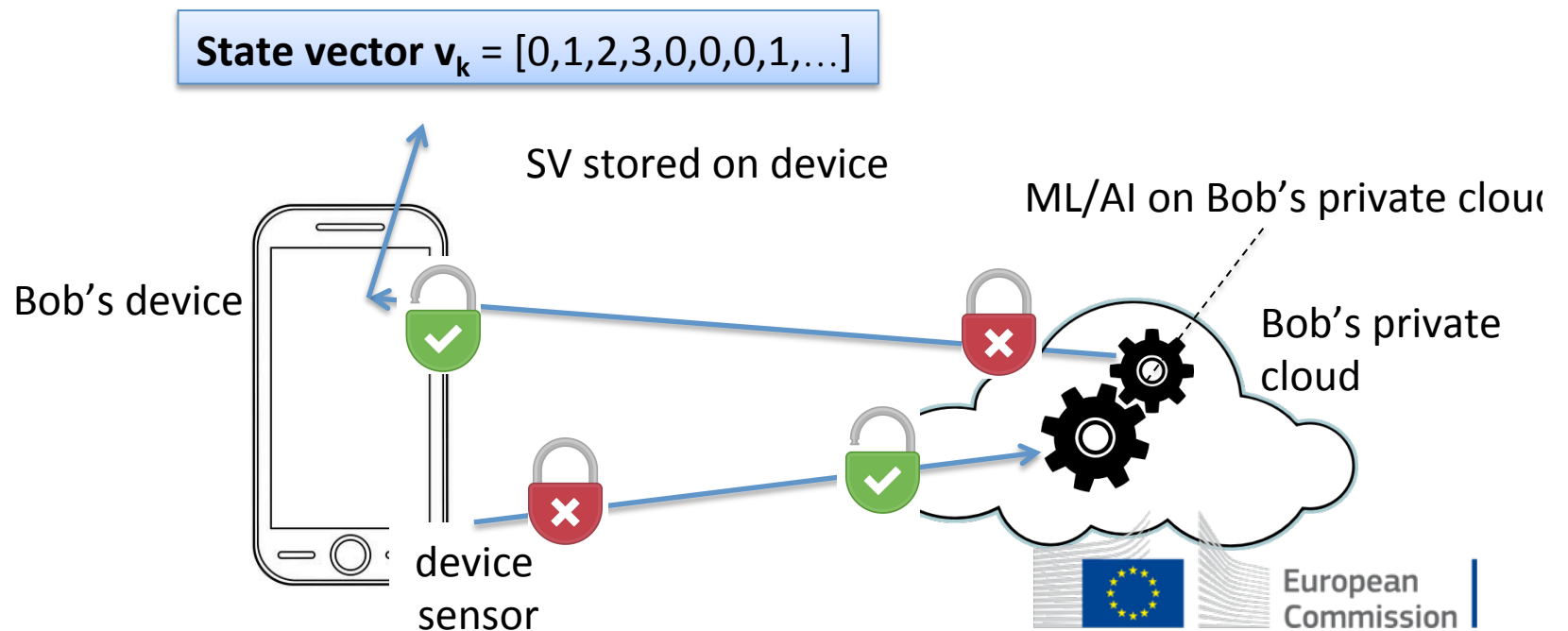
Let's now see how we can build a State Vector for user k ...

1. Full on-device computing (Edge Computing)





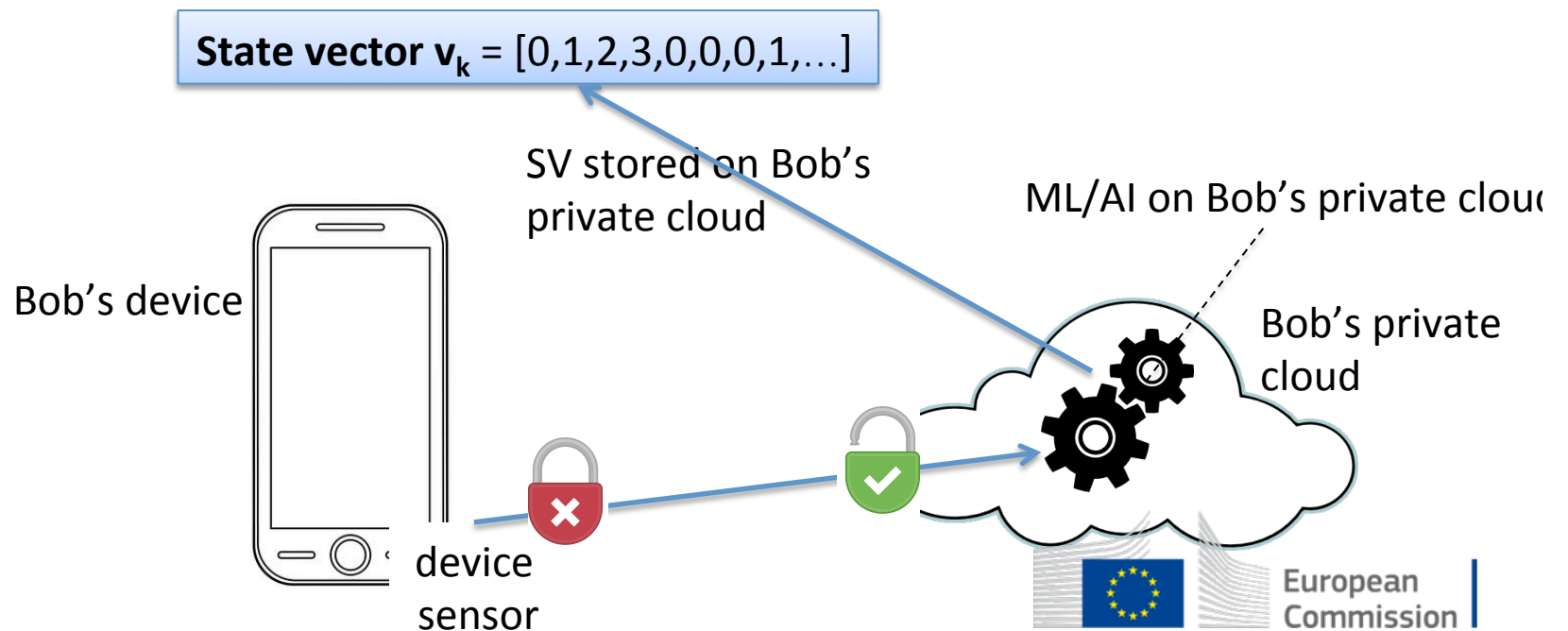
Let's now see how we can build a State Vector for user k ...

1. Full on-device computing (Edge Computing)
2. Support from remote private cloud (device-cloud channel is encrypted,  but data are decrypted  before processing)




Let's now see how we can build a State Vector for user k ...

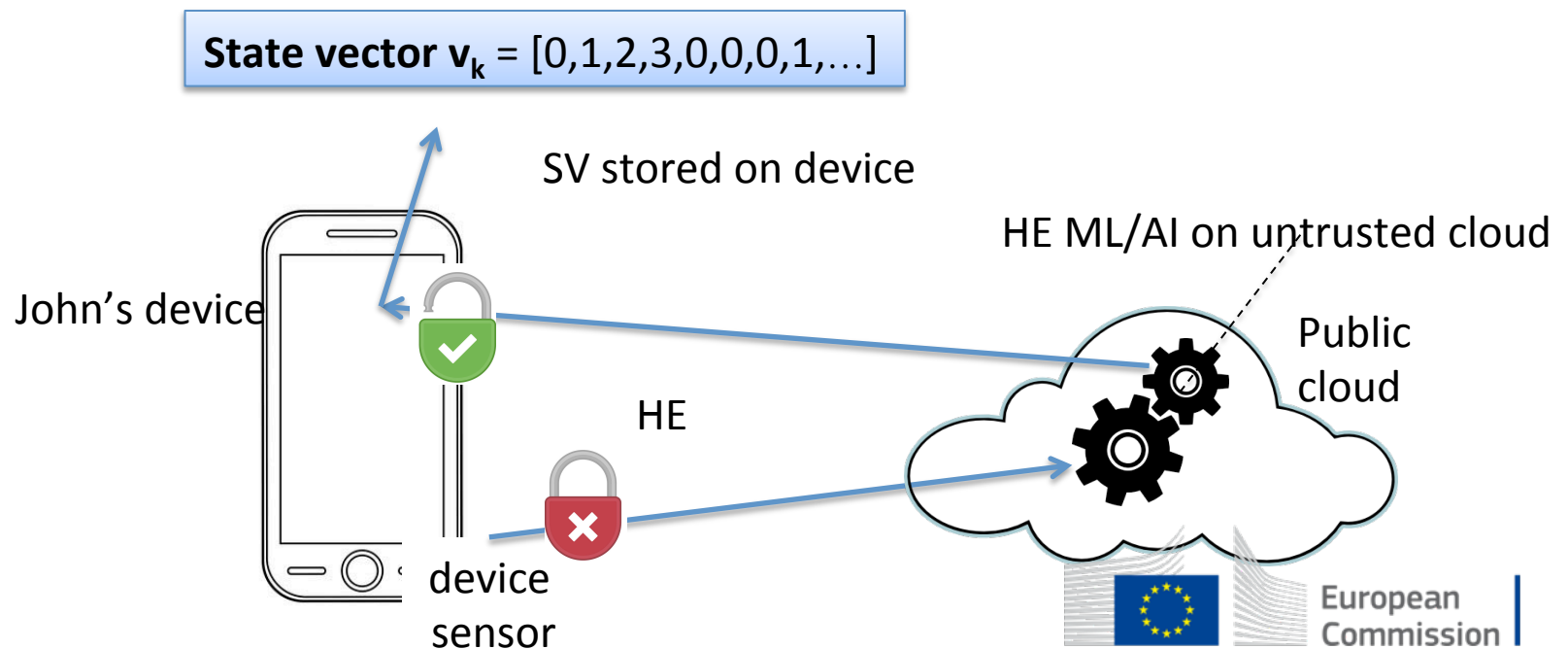
1. Full on-device computing (Edge Computing)
2. Support from remote private cloud (device-cloud channel is encrypted,  but data are decrypted  before processing)



Let's now see how we can build a State Vector for user k ...

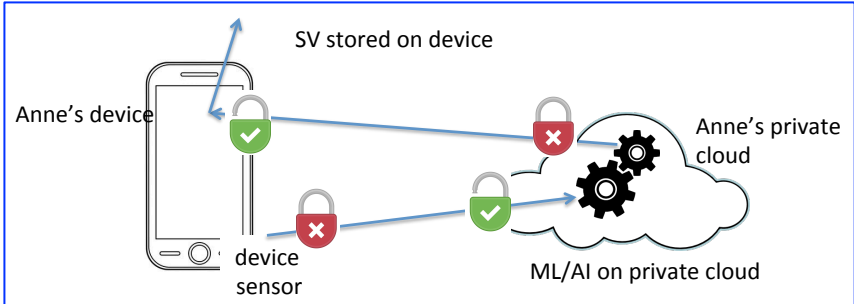
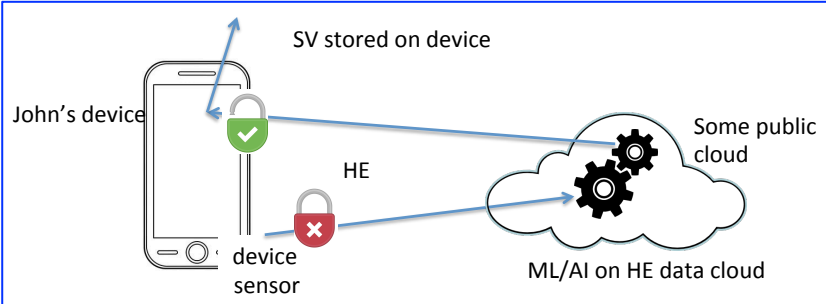
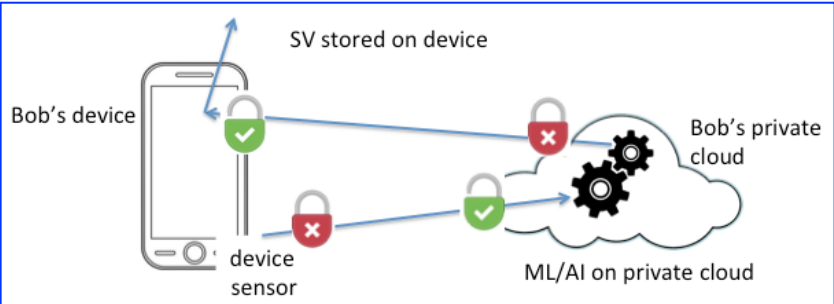
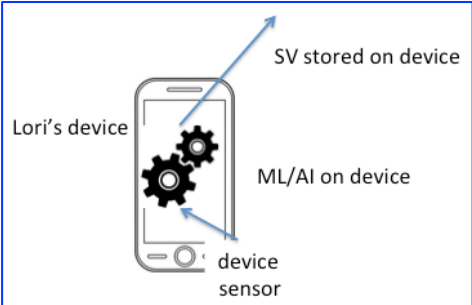
1. Full on-device computing (Edge Computing)
2. Support from remote private cloud
3. Support from remote public with HE data
(variant of previous point: the data are encrypted  with Homomorphic Encryption scheme and stay encrypted during processing and only the results are finally decrypted)

*NB: the HE key for user k is different from the HE key of other users!!!
So in effect, this is just a way to emulate a (single!) private cloud over a public untrusted cloud!! Not much difference from option 2*



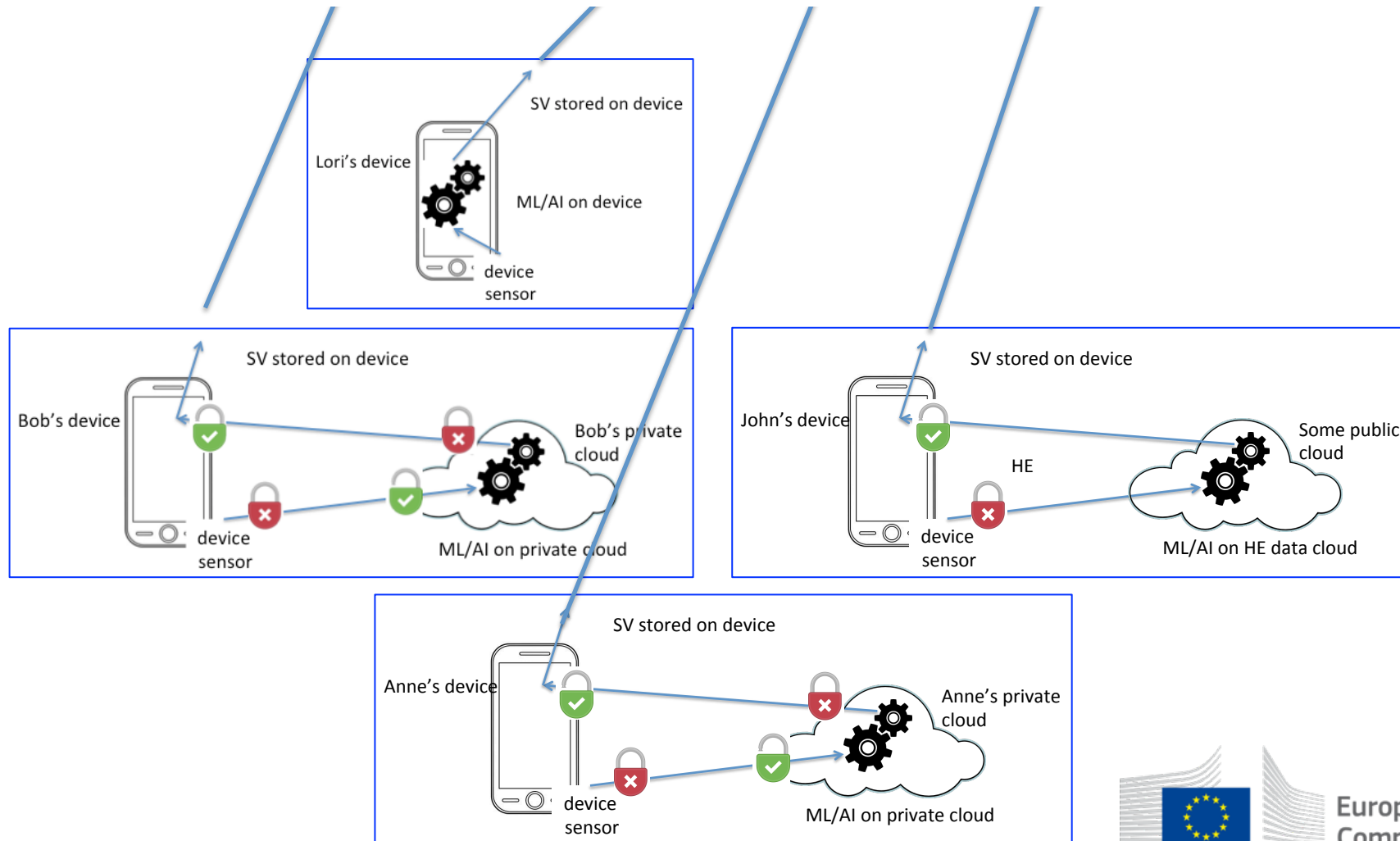
Important: all LP instances for the different users are completely separate from each other!

NB: even if two users use the same (untrusted) public cloud with same HE scheme, they will have different keys! In this way their LP instances remain separate from each other and logically contained within the “private (logical) spaces” of their respective users !!



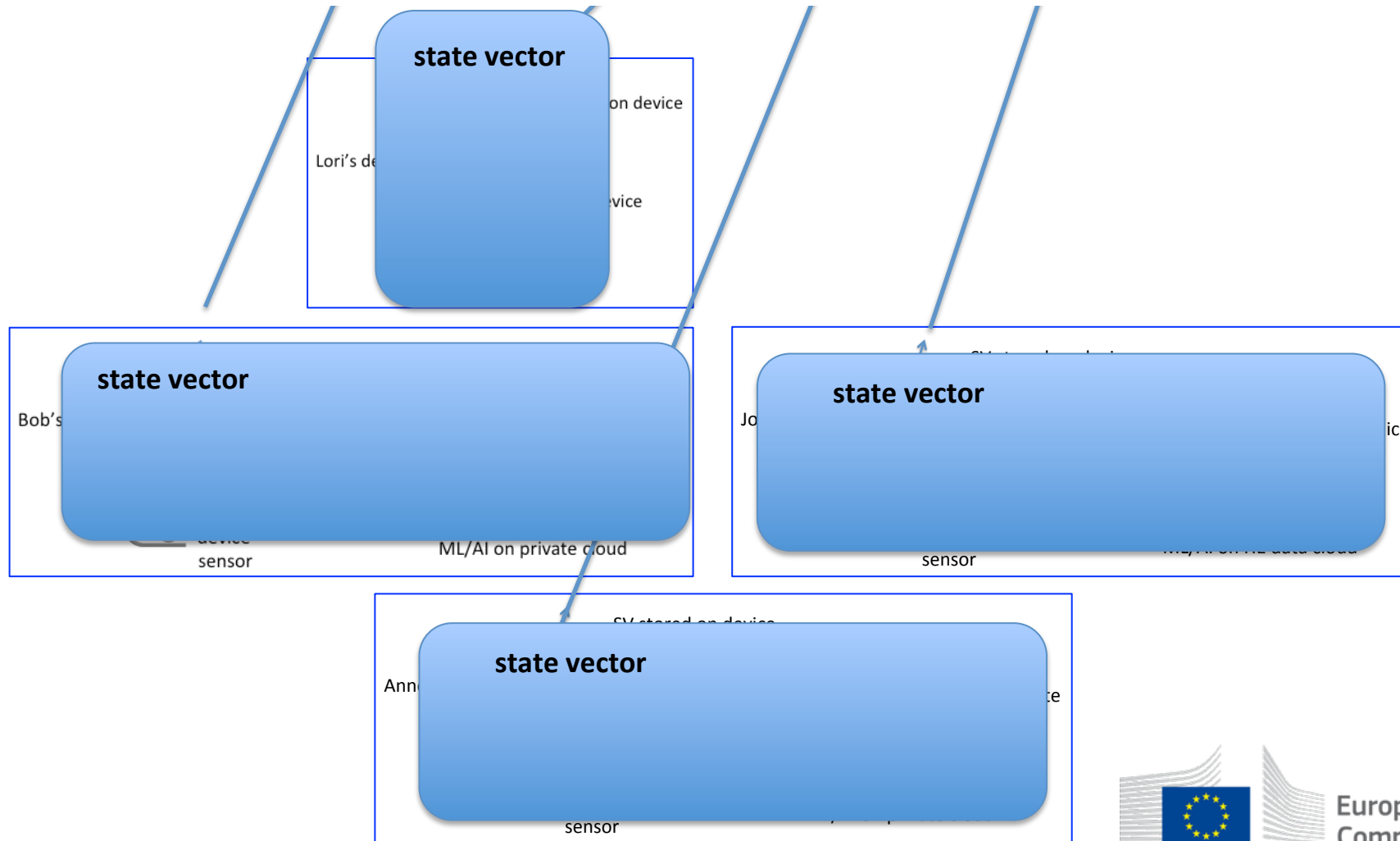
Secure Private Computation Infrastructure

Aggregation over the SV of different users is done only by the GP infrastructure



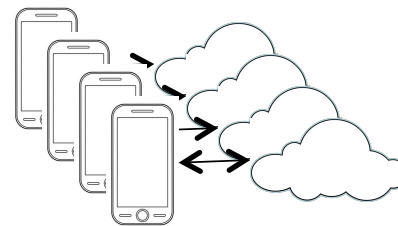
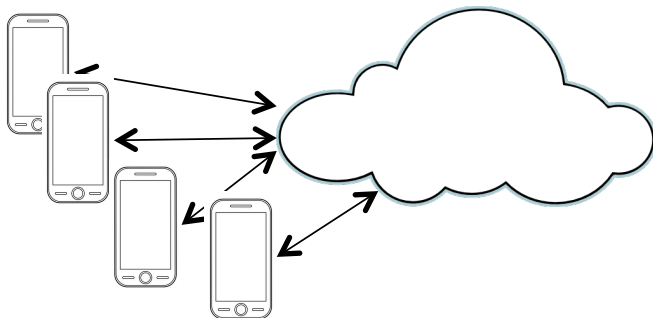
Secure Private Computation Infrastructure

Aggregation over the SV of different users is done only by the GP infrastructure



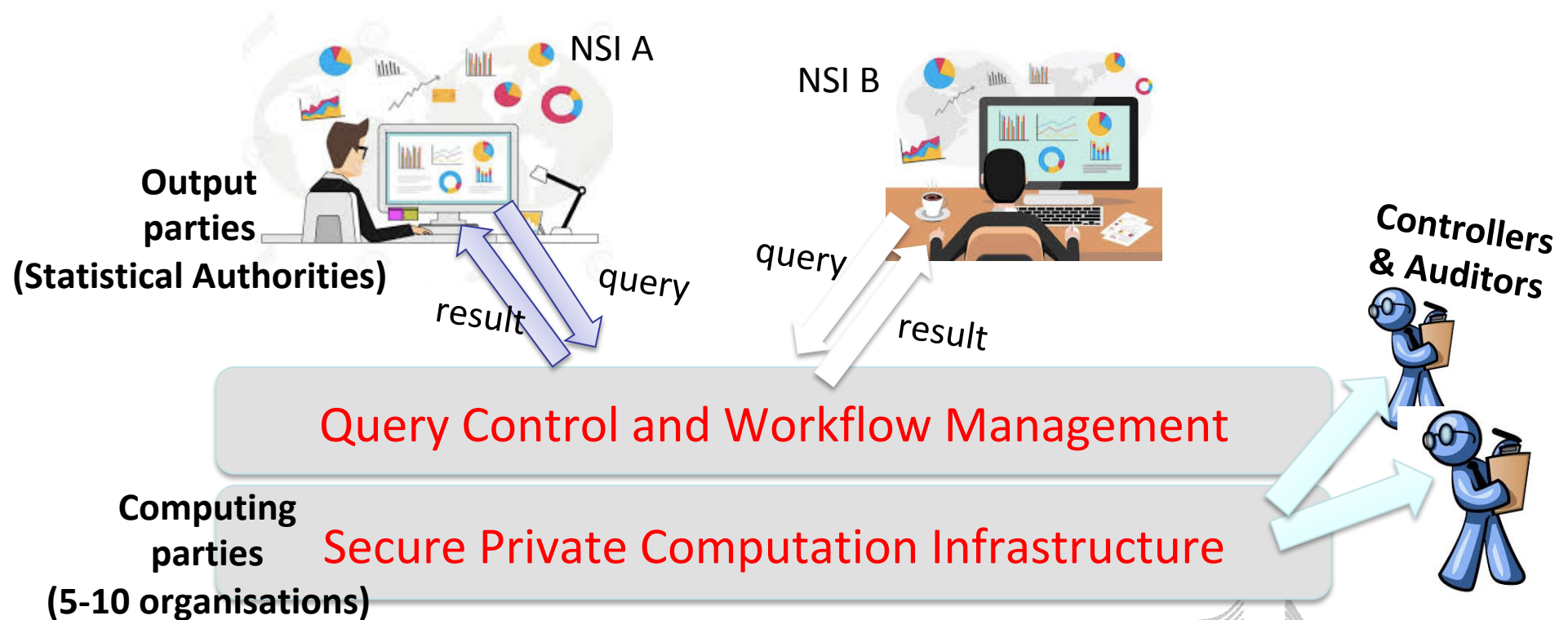
How to reuse an existing Smart Survey tools for building a Trusted Smart Surveys system?

- Say you have a current tool based on a traditional client-server scheme
 - N mobile clients (front-end) + 1 server (back-end) serving all clients
- ... with a single back-end server performing together:
 - 1) the LP (e.g. ML/AI on individual user data) for all N clients
 - 2) storage of LP results, i.e. the SV (State Vector) for all N clients
 - 3) the GP (e.g. clear-text aggregation over all individual SV)
- In such condition, the migration/adaptation scenario would be
 - the LP and GP functions implemented in the back-end must be separated
 - the (legacy) GP function will be replaced by the Secure Private Computation infrastructure (with whatever Secure Private Computing technology, e.g. SMPC, TEE, etc.)
 - the (legacy) LP function can be reused but it will be executed in a private cloud instance for each user, separate from the private clouds of other users



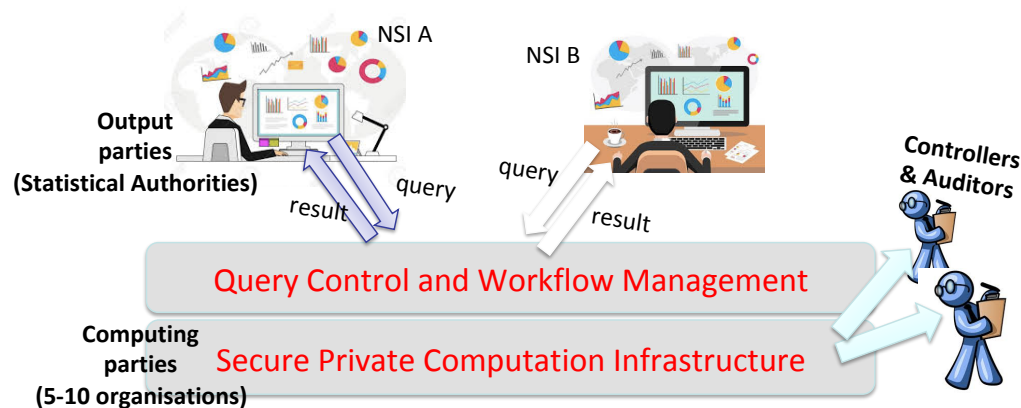
Further thoughts on the Secure Private Computation Infrastructure

- Initial focus of the design should be on simple GP functions (e.g. additive operations, simple primitives ... but not simplistic)*



Further thoughts on the Secure Private Computation Infrastructure

- *Initial focus of the design should be on **simple GP functions** (e.g. additive operations, linear regression, simple primitives...)*
 - **Computationally heavy operation are ok in the LP, but may not be supported by (the first version of) GP.**
- *By keeping **GP operations simple** (in the initial design) → we don't have to worry much about **scalability** → therefore we can concentrate the specification efforts towards other strategic dimensions, like: stronger **security model** (redundancy of ex-ante and ex-post checks), **verifiability** and robustness against active attacks, **multi-vendor compatibility** (produce standards!)*



Follow-up

Email: Fabio.Ricciato@ec.europa.eu

References

Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics (short paper for SIS 2020),
https://ec.europa.eu/eurostat/cros/content/trusted-smart-surveys-possible-application-privacy-enhancing-technologies-official-statistics-short-paper-sis-2020_en

Trusted smart statistics: Motivations and principles (SJIAOS 2020)
https://ec.europa.eu/eurostat/cros/content/trusted-smart-statistics-motivations-and-principles_en