The role of Privacy Enhancing Technologies in future Trusted Smart Statistics

Fabio Ricciato,
Eurostat, Unit B1 ‘Methodology & Innovation in Official Statistics’

fabio.ricciato@ec.europa.eu

CYD Conference on PET
4· November 2020

What is Eurostat?

Part of the European Commission

The statistical office of the European Union

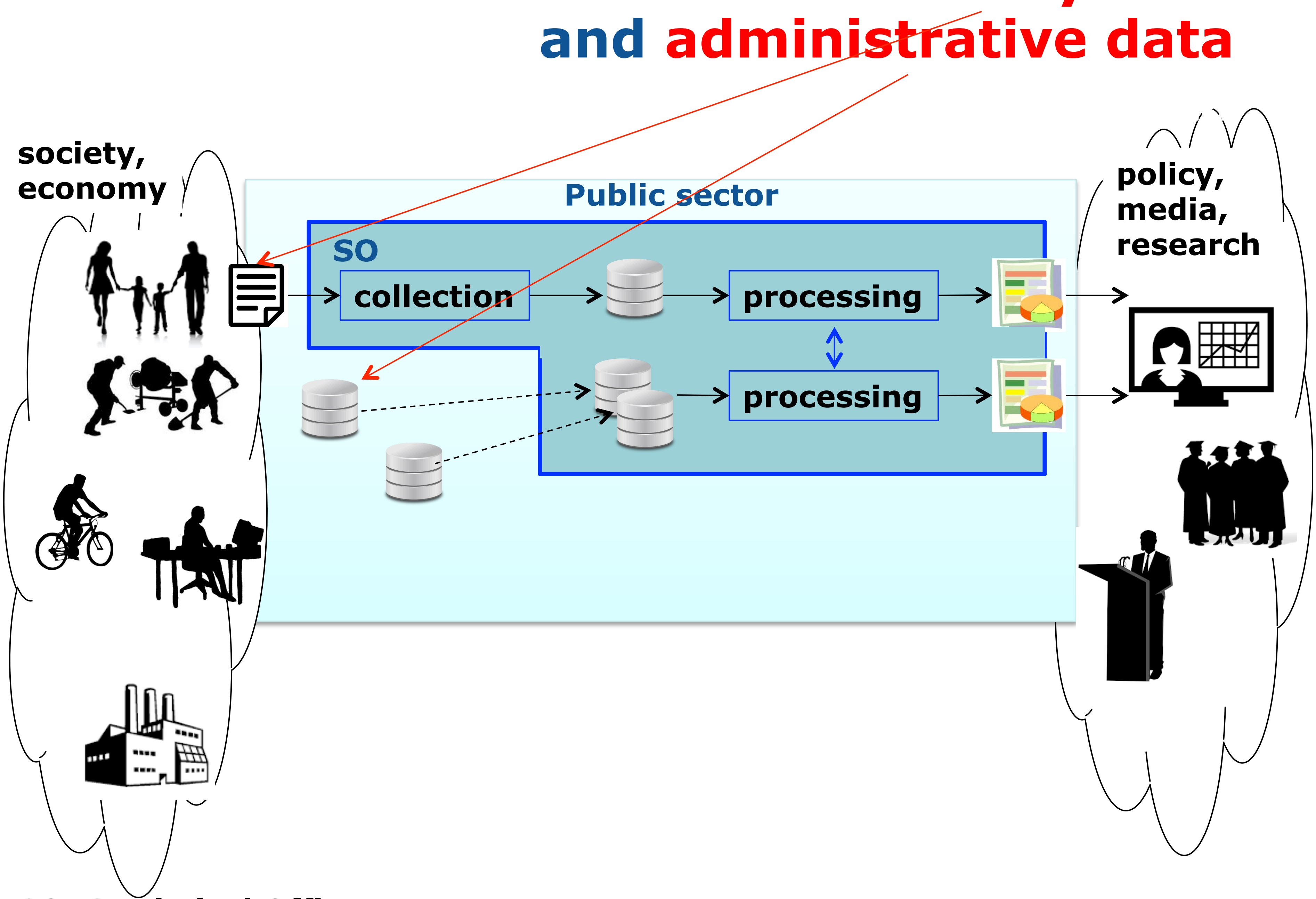
The central institution of the European Statistical System (ESS)

Official Statistics

The role of official statistics is to describe quantitatively the economy, the society and the environment

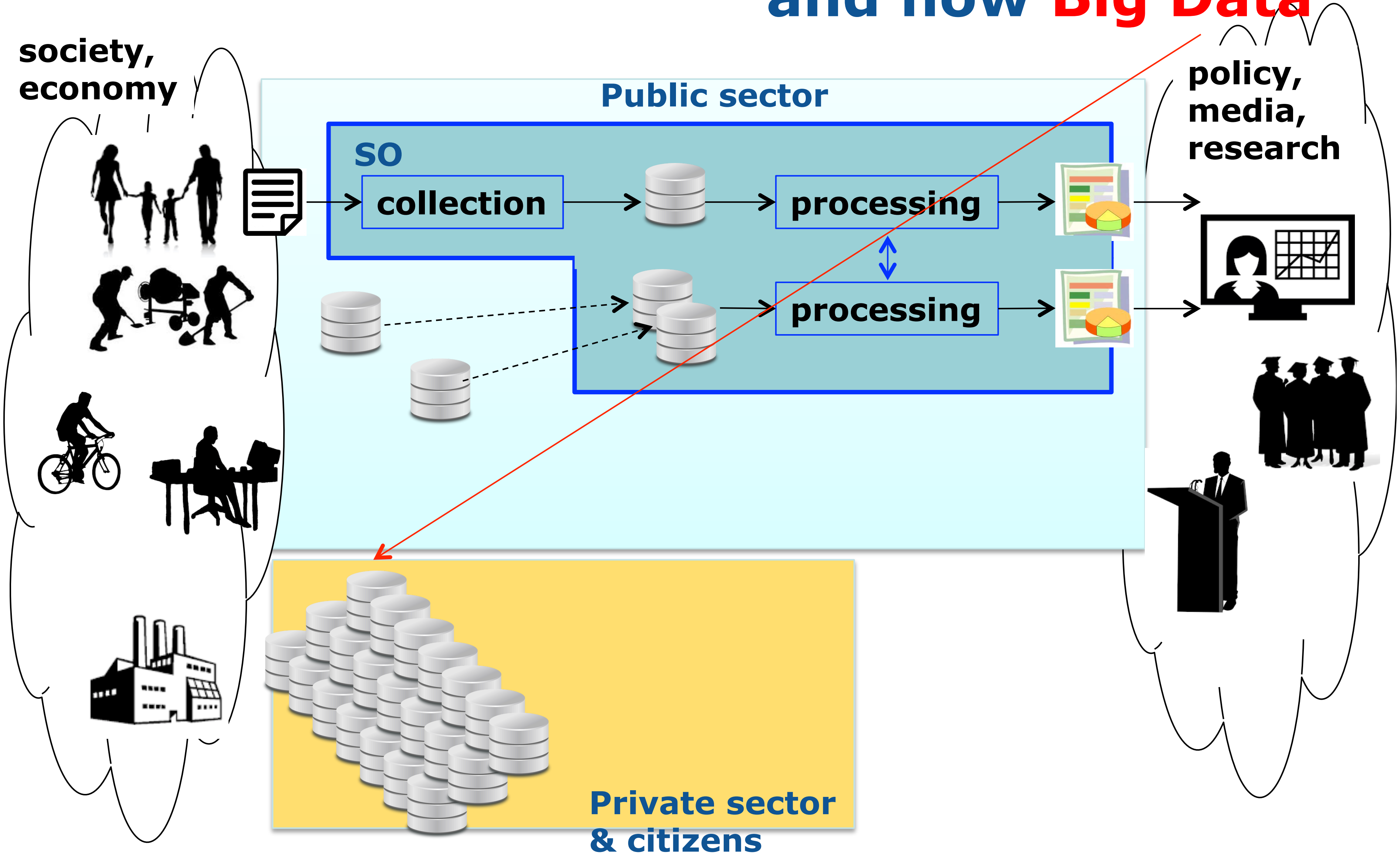
European Commission

Official Statistics based on survey data and administrative data



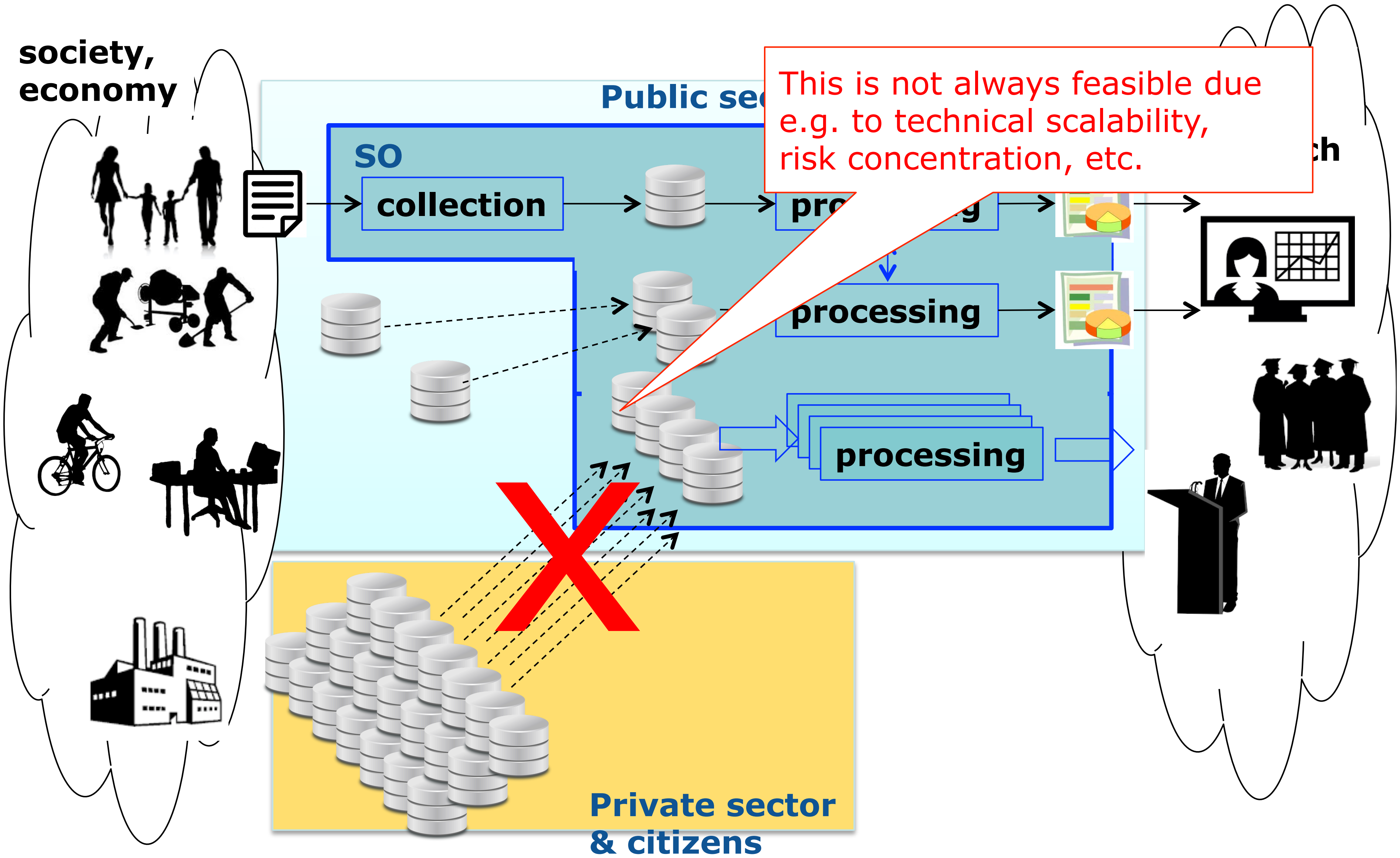
SO: Statistical Office

Official Statistics based on survey data and administrative data and now Big Data



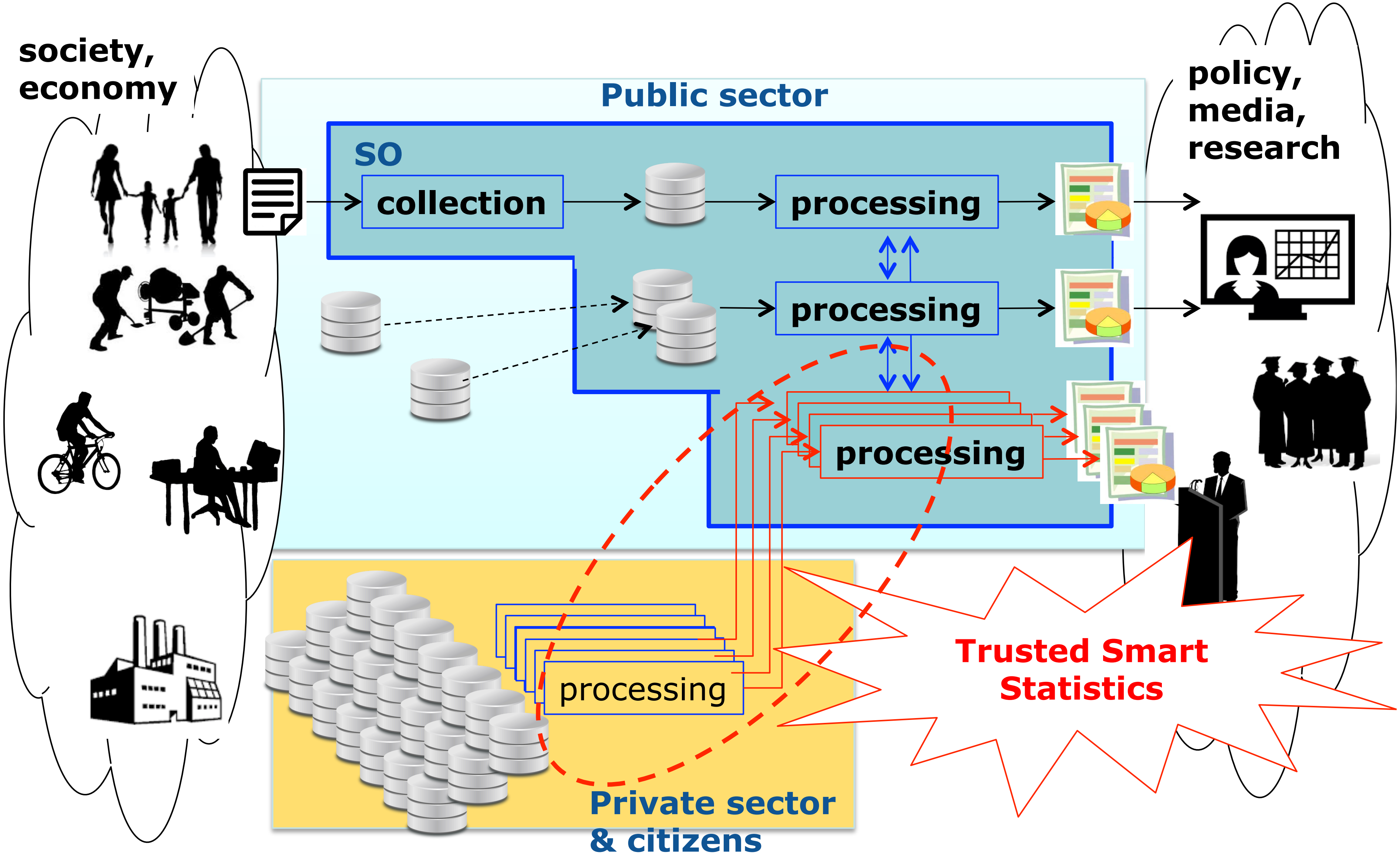
Handling the new in the old way

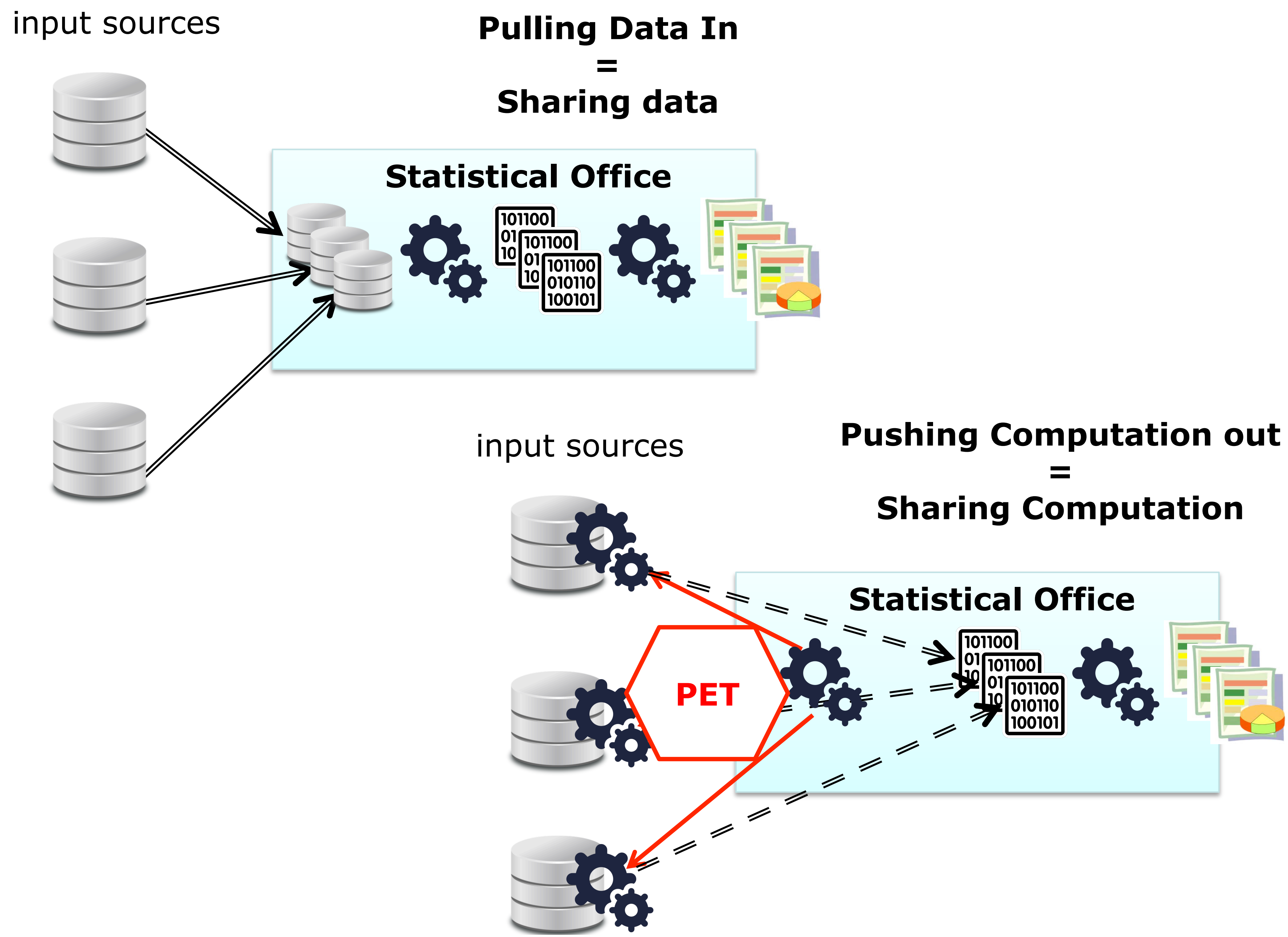
Pull data in



Handle the new in new ways

Push computation out (partially)





PET: Privacy Enhancing Technologies

Data and new data

"micro-data"

Name. Gender. Birth date.

Marital Status. Residence address.
Occupation. Household composition…
…
Monthly income.
Monthly expenditures per good category.
Number of touristic trips in a year
…

"nano-data"

…

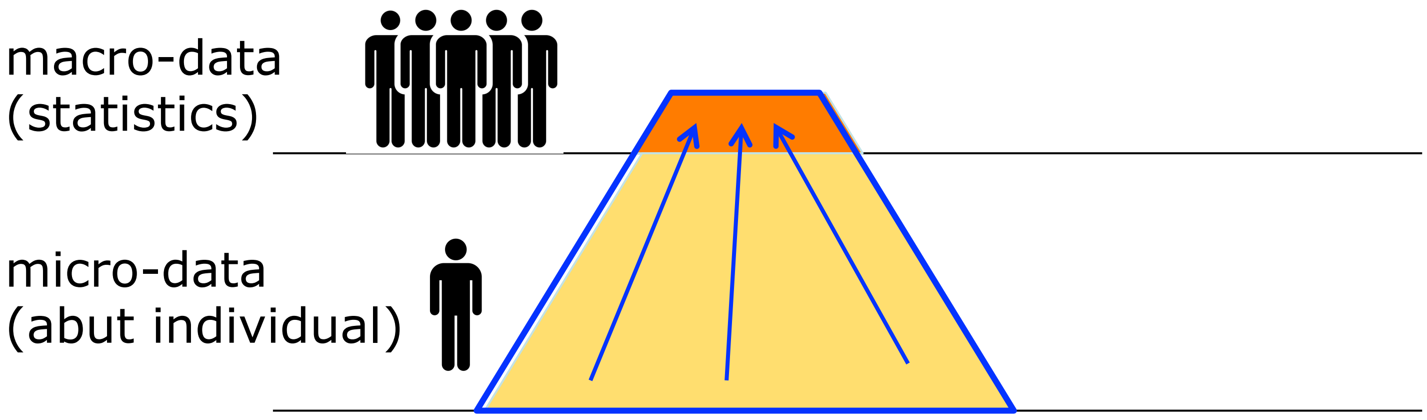
Your exact location, every second.
Every single heart-beat, blood pressure…
Every single transaction, purchases, encounter, event involving you…
Your current opinion on any single fact…

"Deep data"

European Commission

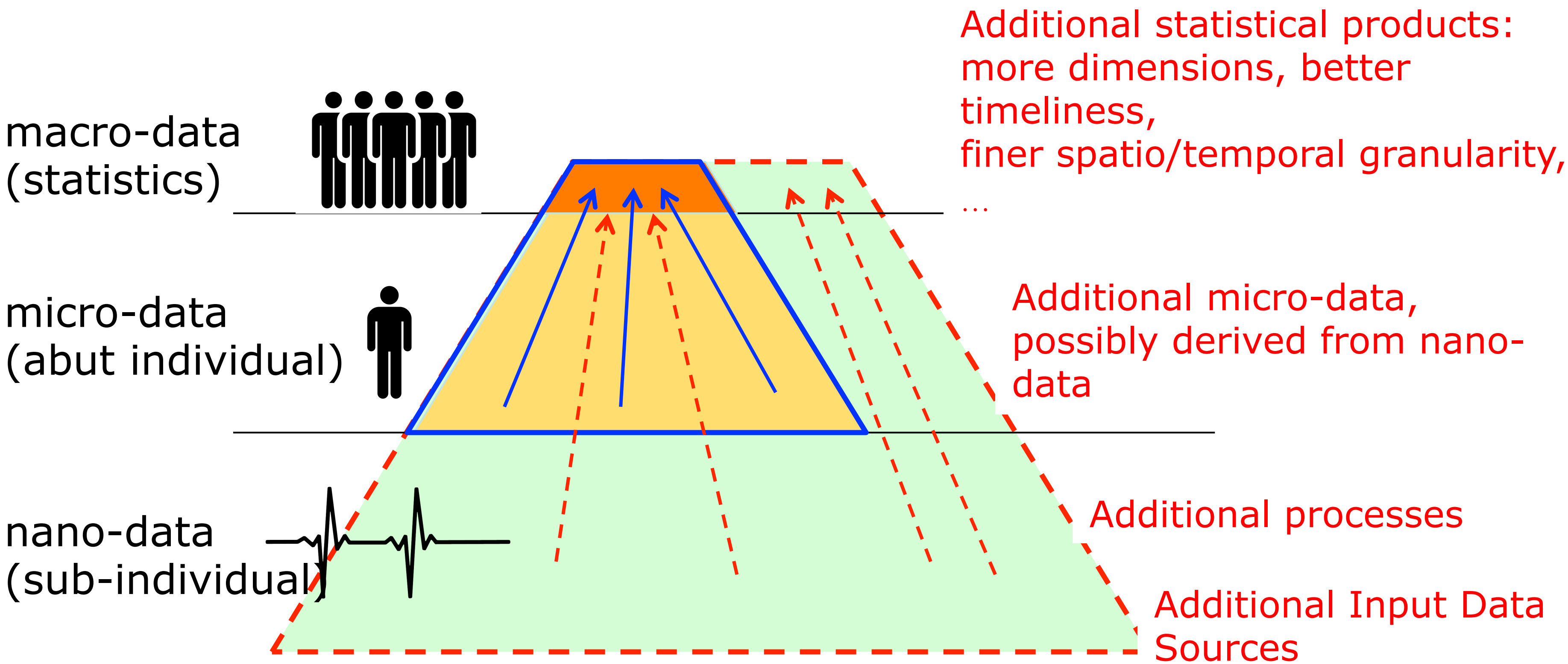
Official Statistics.

- *The ultimate goal of Official Statistics is to produce macro-data (statistics) from input micro-data*
  - **Collection of micro-data as ancillary task**



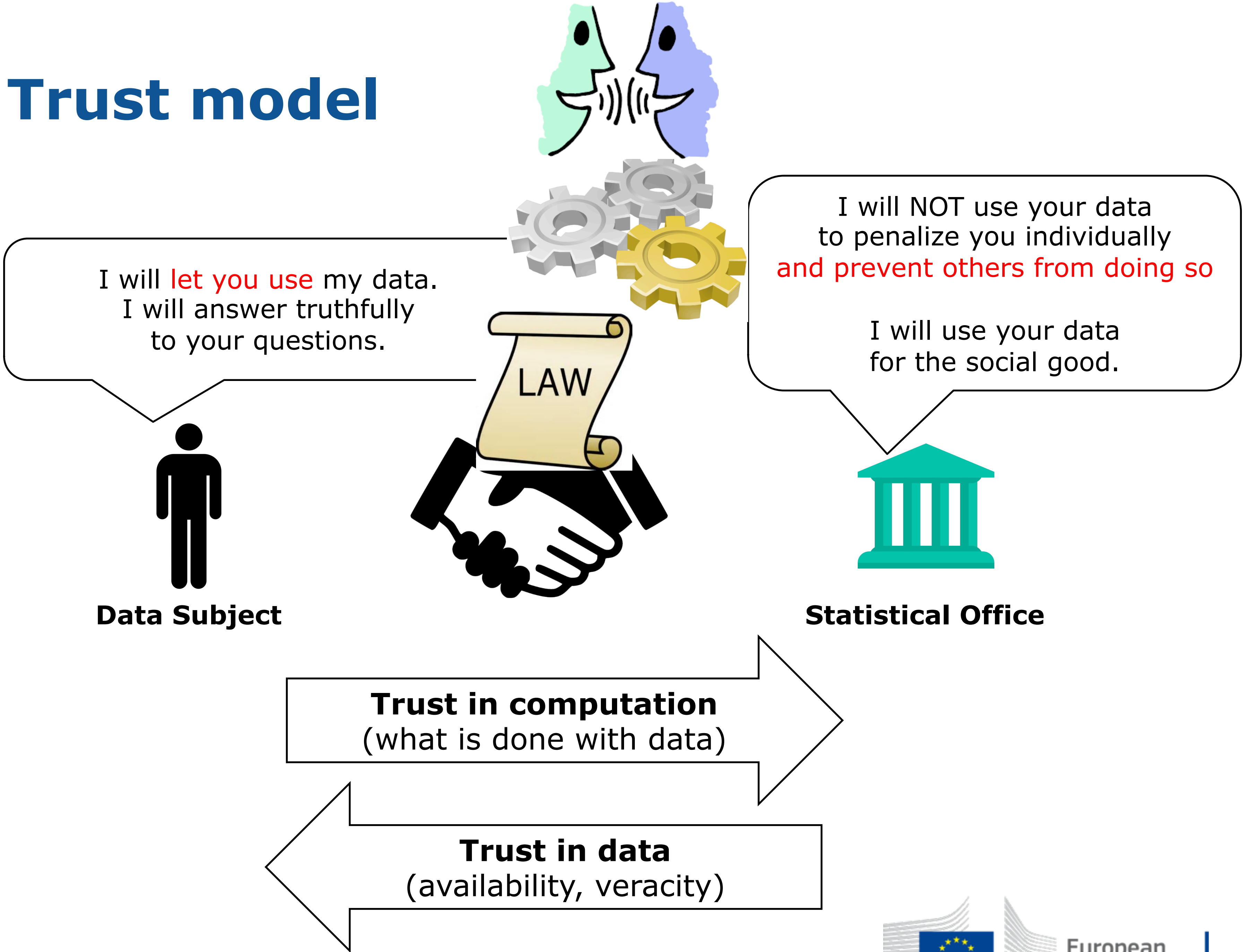
Official Statistics. Augmented

- *Availability of new (deep, nano) data sources as opportunity to extend & empower Official Statistics*



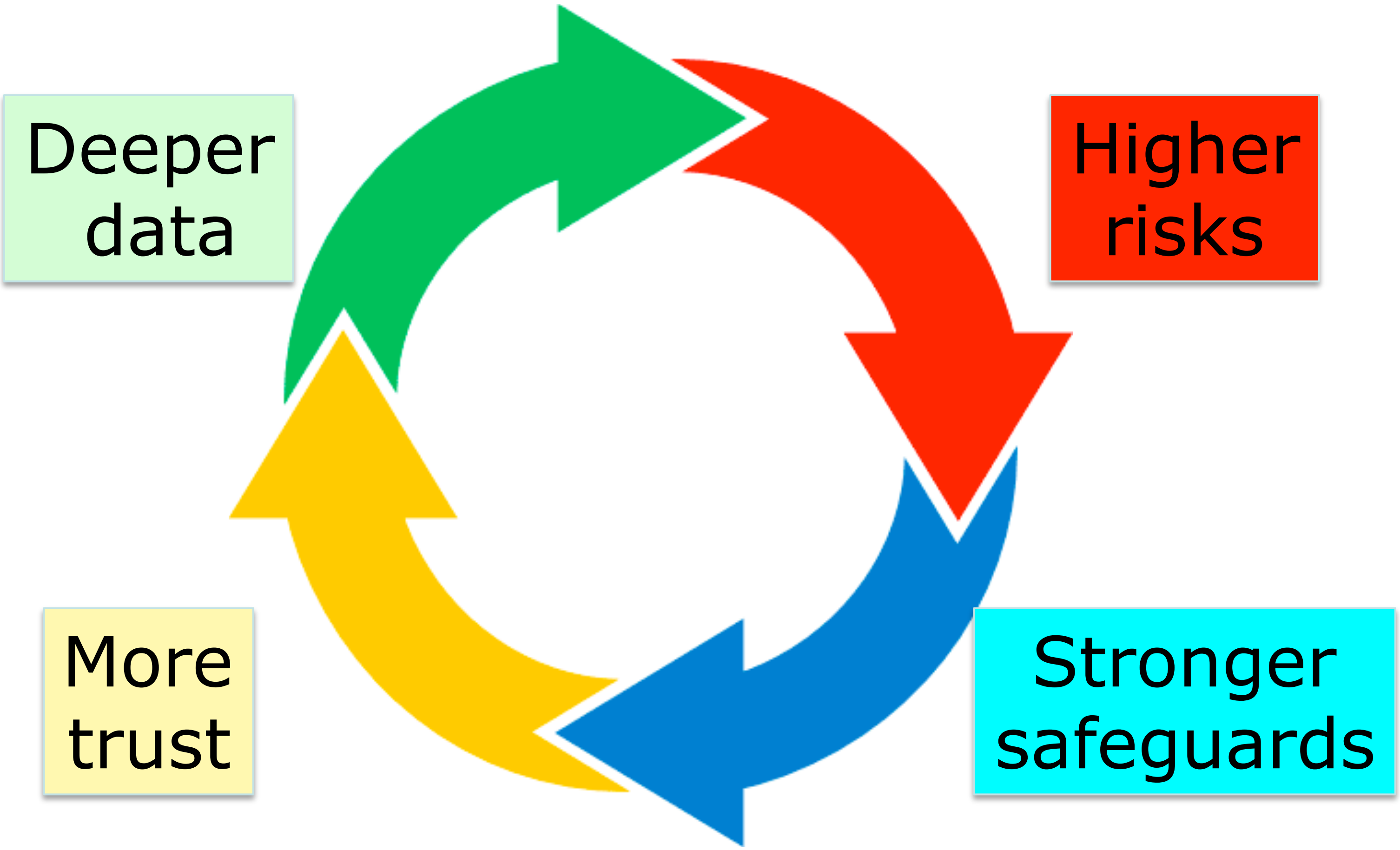
European Commission

Trust model



European Commission

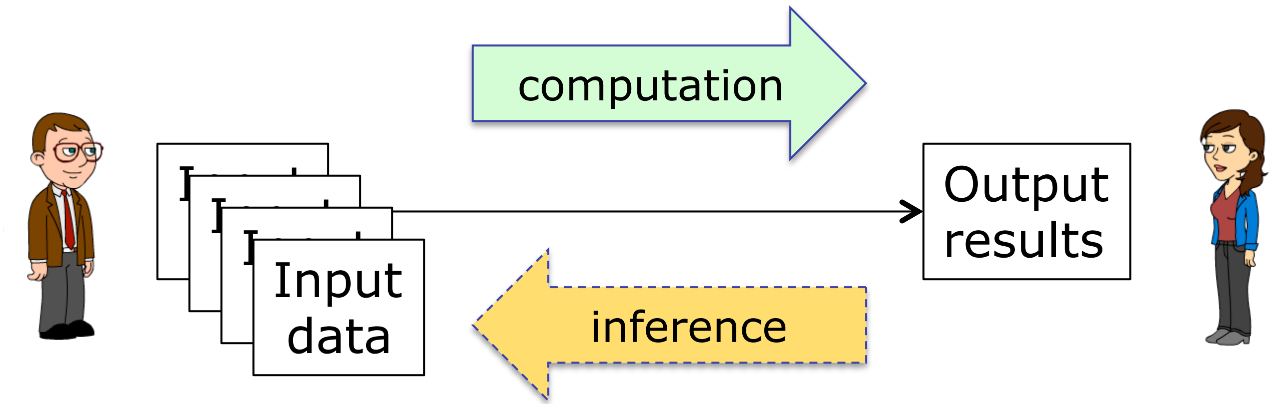
Smart & Trusted



- *Input Privacy vs. Output Privacy*

TEE HE SMC

Input privacy problem: **enabling forward computation**
(from closed input)



Output privacy problem: **preventing backwards inference**
(from disclosed output)

SDC DP

SMC: Secure Multi-party Computation
SDC: Statistical Disclosure Control

TEE: Trusted Execution Environment
HE: Homomorphic Encryption
DP: Differential Privacy

Output Privacy:
Statistical Disclosure Control (SDC)

- *Suppression (e.g. cell deletion, column removal)*
- *Add noise, perturbation, rounding*

Town	Count all	Count sick	Average Income
…	…	…	…
Smallville	5	1	51
Midpoli	85	7	40678
Largetown	5777	45	89

SDC

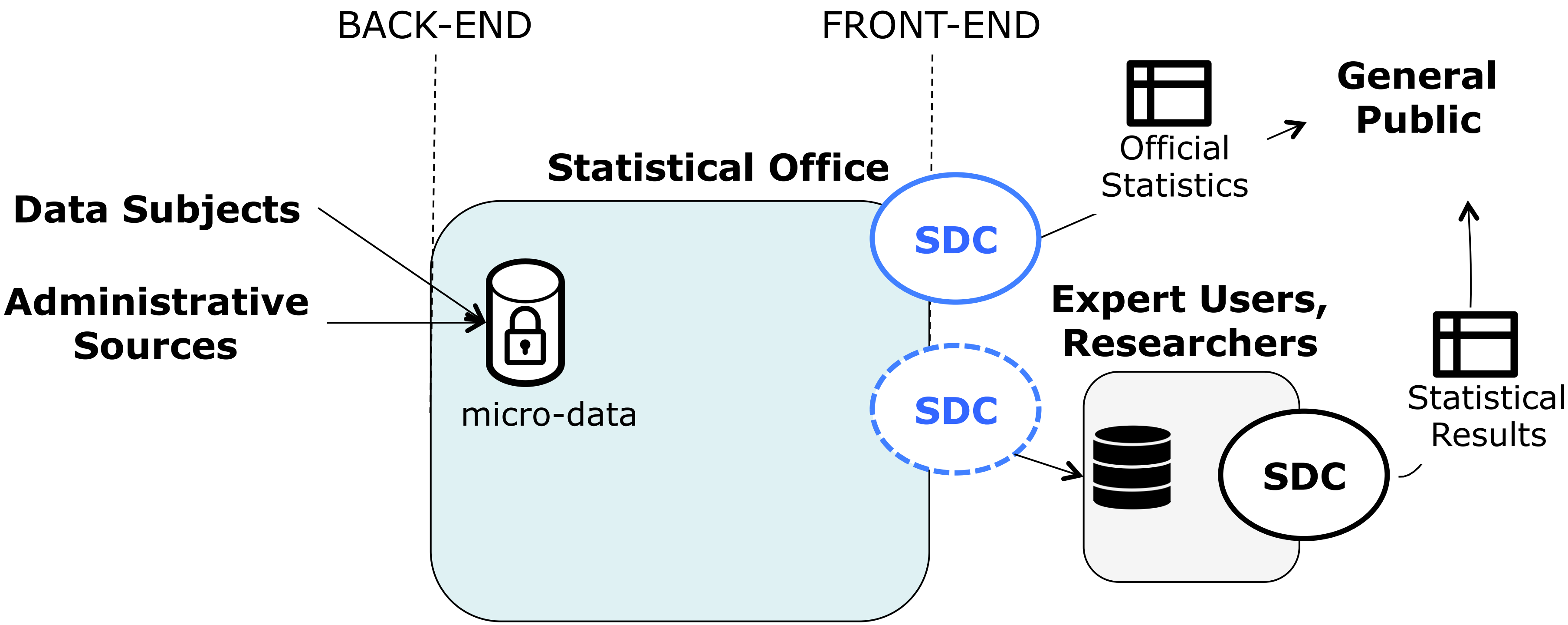
Town	Count all	Count sick	Average Income
…	…	…	…
Smallville	6	2	59
Midpoli	88	7	40401
Largetown	5773	44	89

Name	Age	Gender	Income	Town	Sick	
…	…	…	…	…	…	
Eva	23	F	10	Smallville	1	
Fabio	38	M	30	Largetown	0	
Elisa	78	F	100	Largetown	1	
Oscar	32	M	23	Midpoli	0	
Michail	38	M	40000	Midpoli	0	
Anna	24	F	11	Largetown	0	…
…	…	…	…	…	…	…

SDC

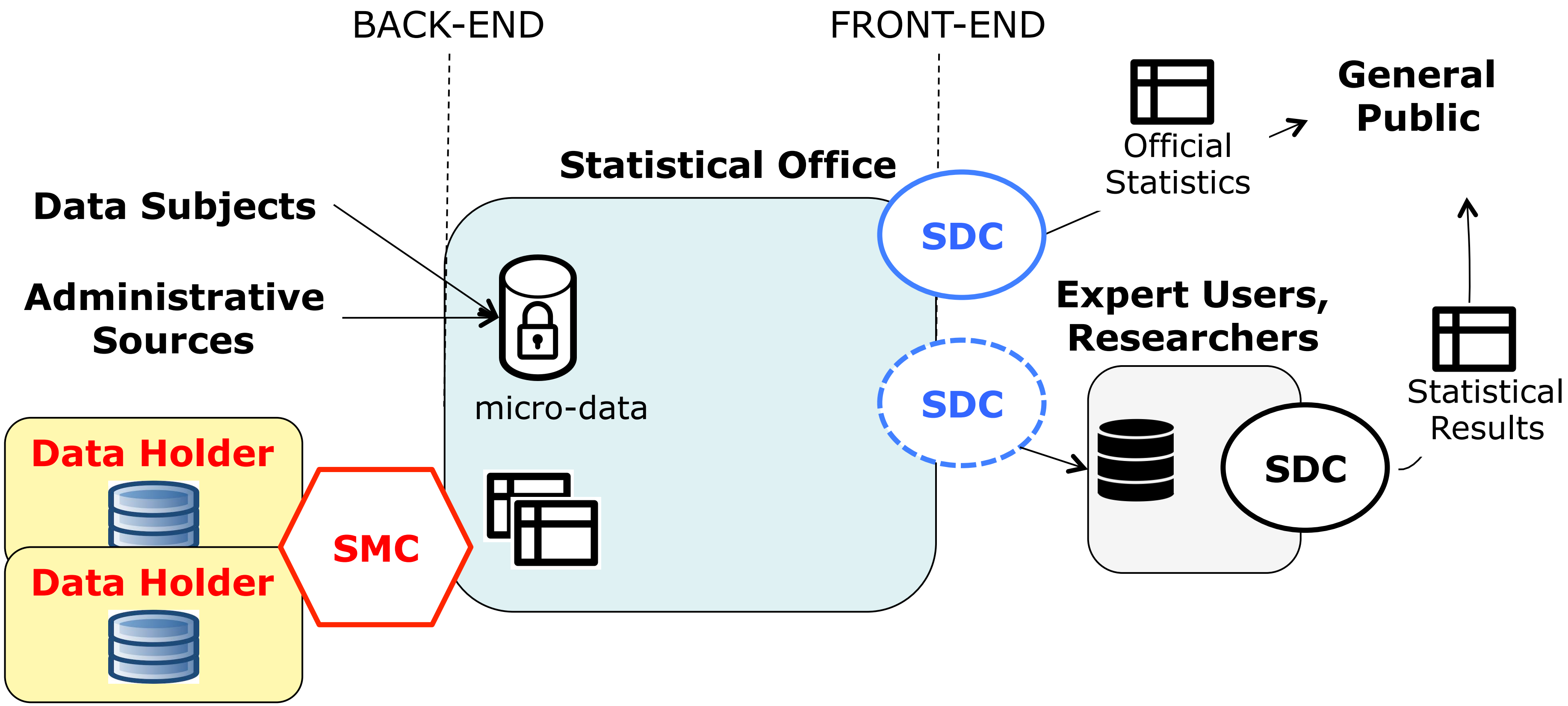
…me	Town	Sick
	…	…
	Largetown	1
11	Largetown	0
100	Smallville	1
23	Midpoli	0
40000	Midpoli	0
30	Largetown	0
…	…	…

SDC on the front-end



SDC: Statistical Disclosure Control

SMC on the back-end



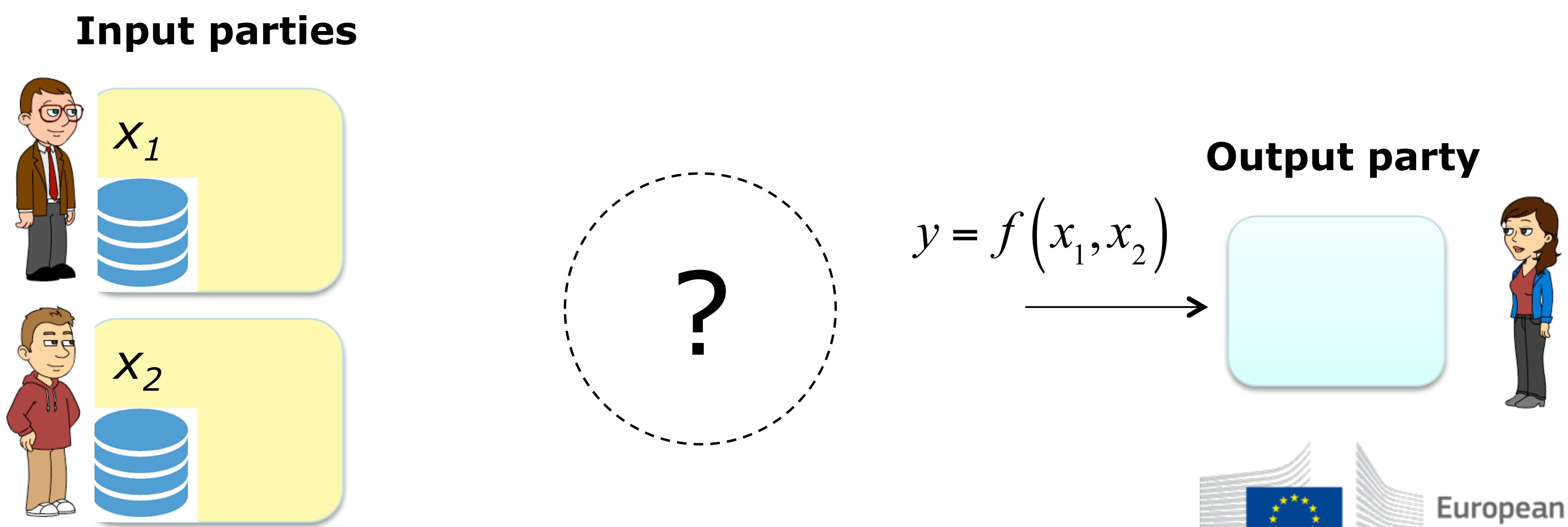
SDC: Statistical Disclosure Control

SMC: Secure Multi-Party Computation

- *Input Privacy approaches for multiple input parties*

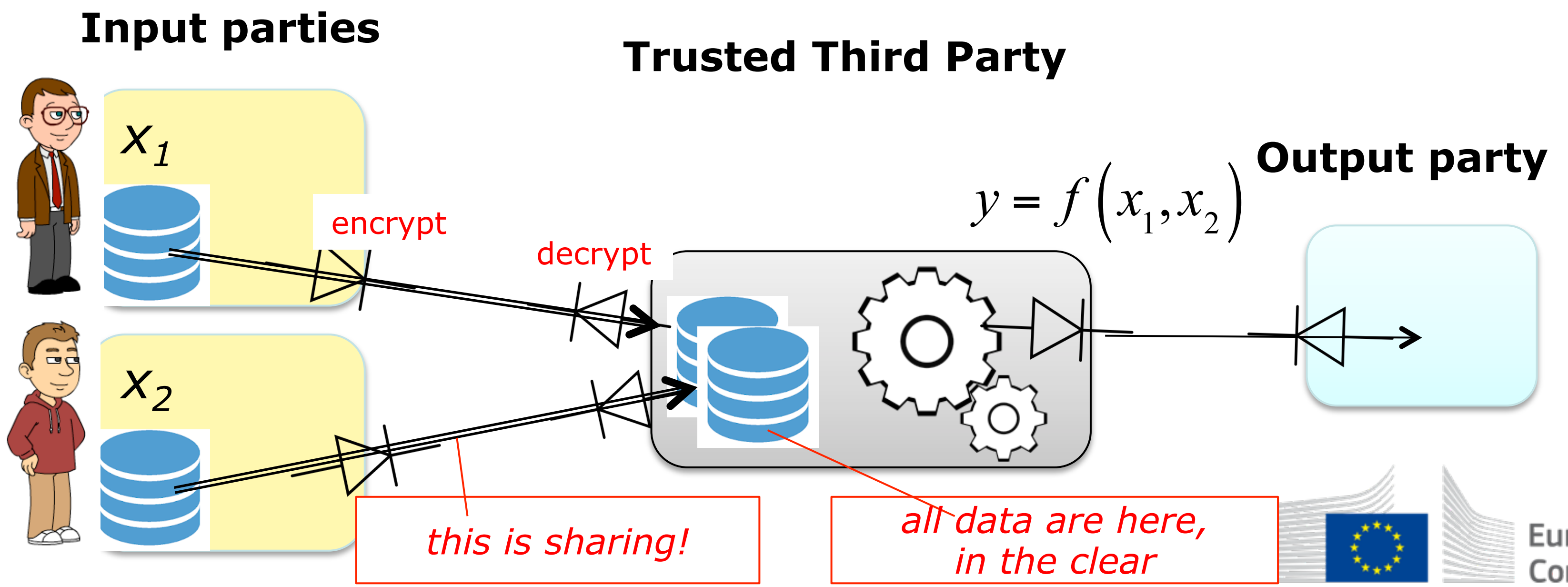
Input Privacy problem

- *Marc and Bob (the input parties) agree to let Anne (output party, or result party) learn the result y=f(x1,x2)*
- *But nobody wants to share their input to any other…*



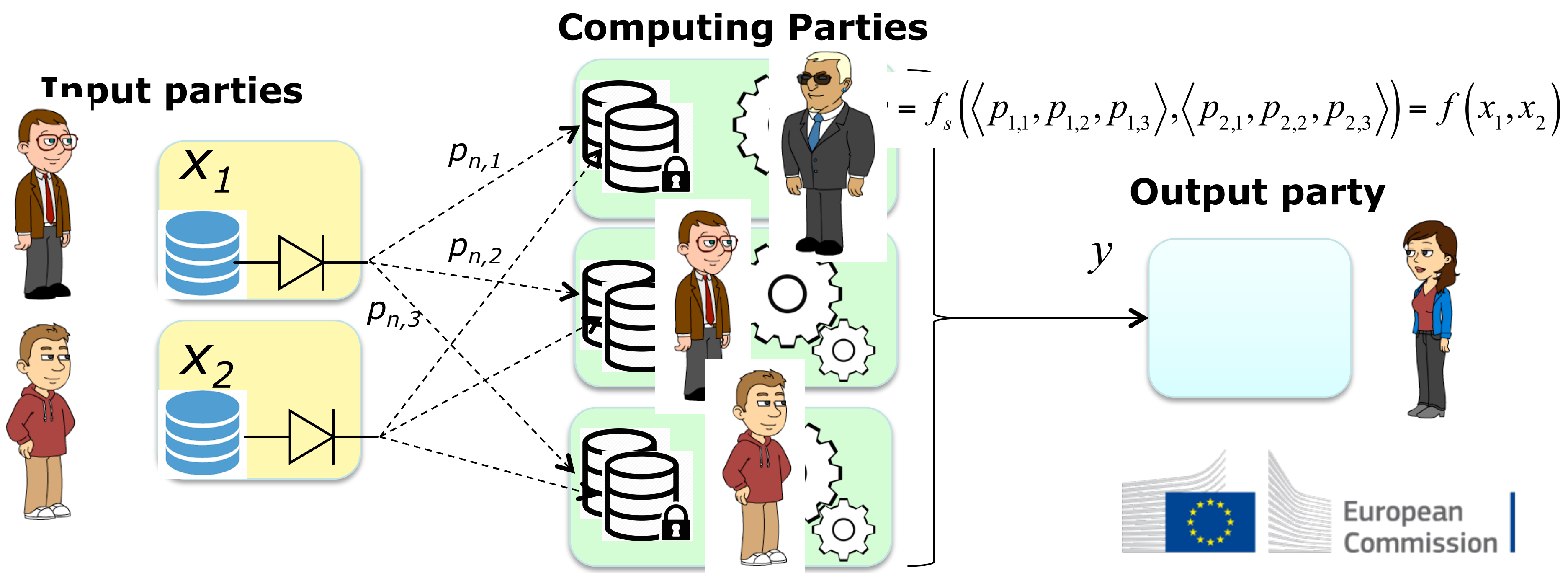
Trusted Third Party (TTP)

- *Wih a Trusted Thid Party (TTP) …*
  - **data sharing still occurs towards the TTP**
  - **risk concentration: TTP gets all the data**
    → single point of (trust) failure
  - **a single entity trusted by all parties might not exist**



Secure Multi-Party Computation (SMC)

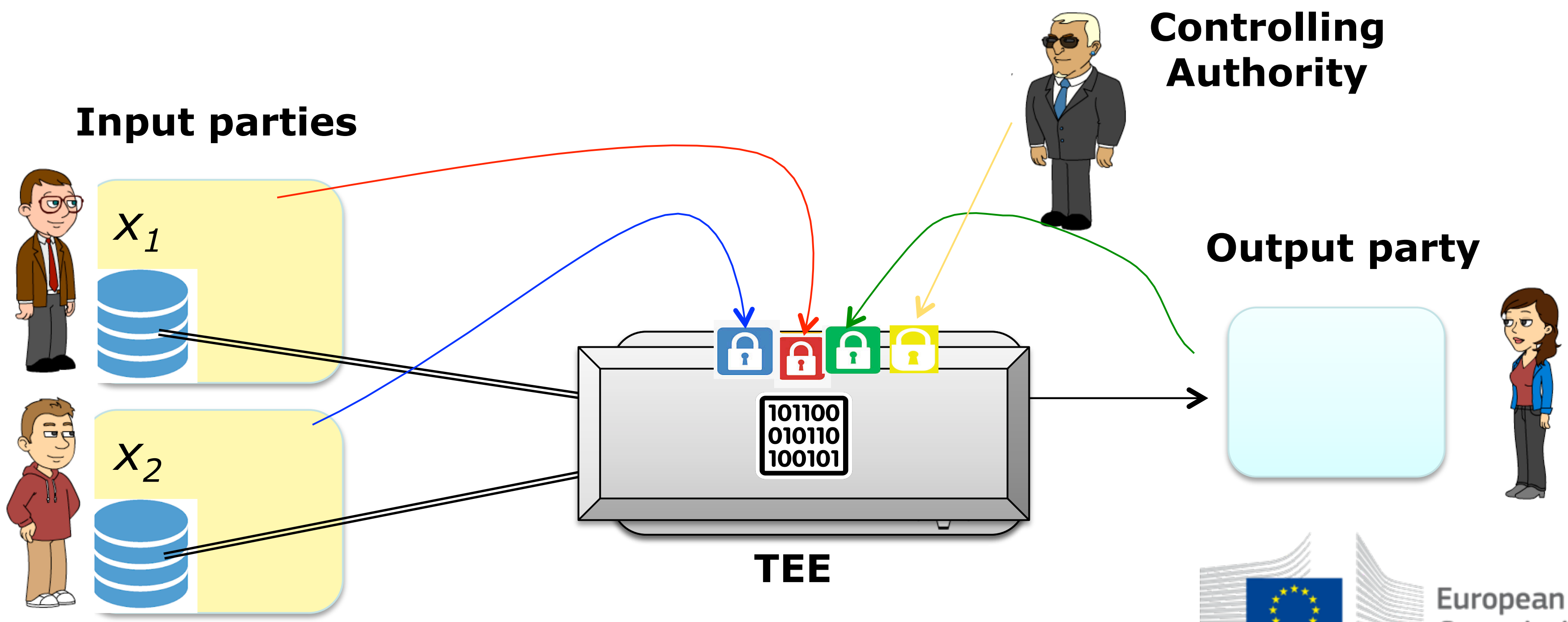
- *Each element of secret input x_n is transformed into K "shares" $p_{n,1}$, $p_{n,2}$ … $p_{n,k}$ that are distributed to different* *computing parties*
  - **no single party holds "the data"**
- *The computation on secret shares*
  - **is distributed (shared) among the computing parties**
  - **returns the same output value that would be obtained from the input data (homomorfism)**
- *The computing parties need to be trusted collectively, not individually*



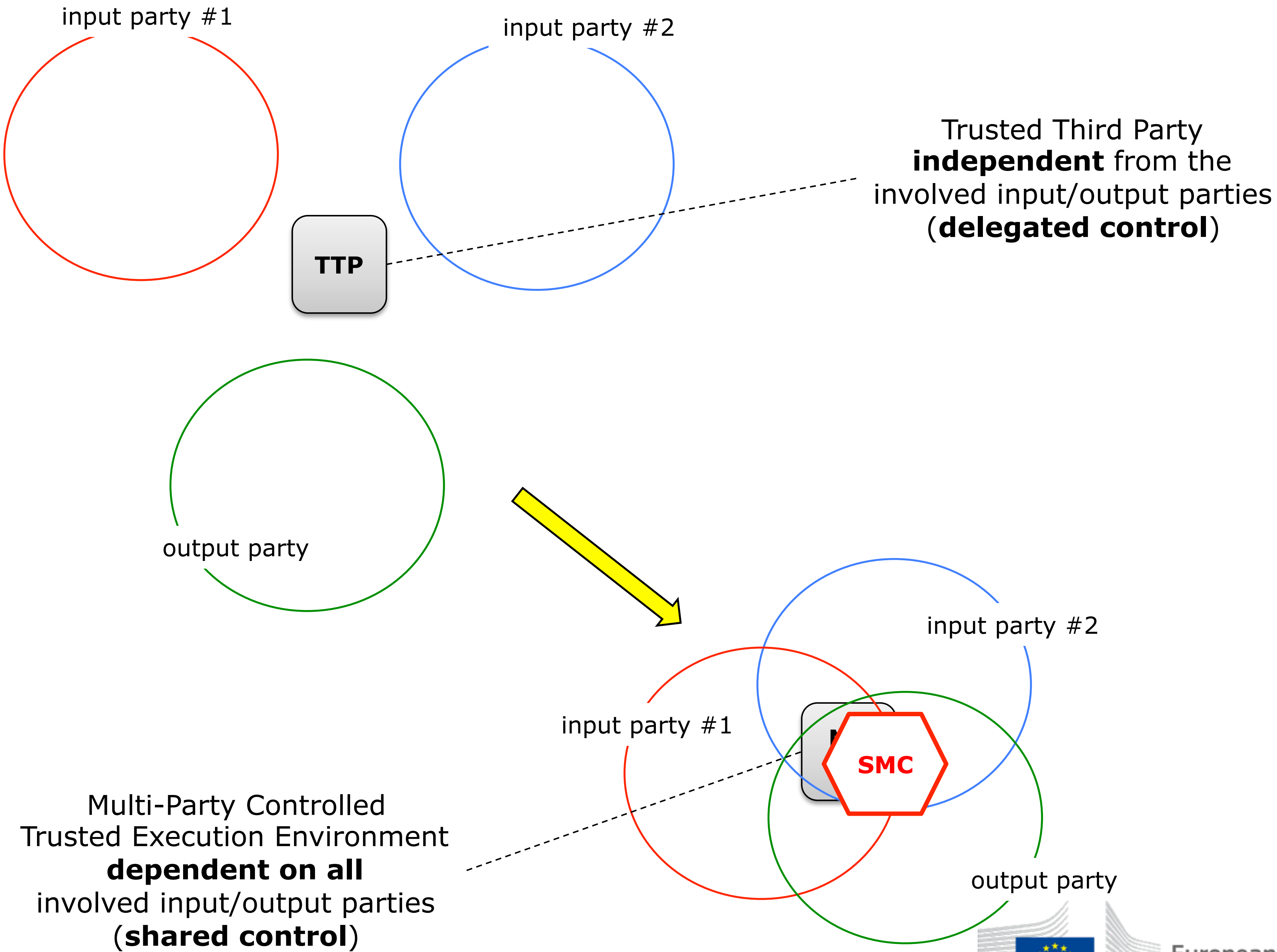
European Commission

Multi-Party Controlled Trusted Execution Environment (MPC-TEE)

- *Think of a special machine, "trusted" **at all layers** (software & <u>hardware</u>) to execute/install only code that was **jointly authenticated by all involved parties***
- *TEE separates "ownership" and "control" of the machine*



European Commission



European Commission

Trust & Control

- Who has control, has to be trusted
- Building trust = engineering credible schemes
  for controlling ***the use of* data**.
- Centralised control → single-point-of-trust
- **Shared control** → trust multiple entities collectively, not individually

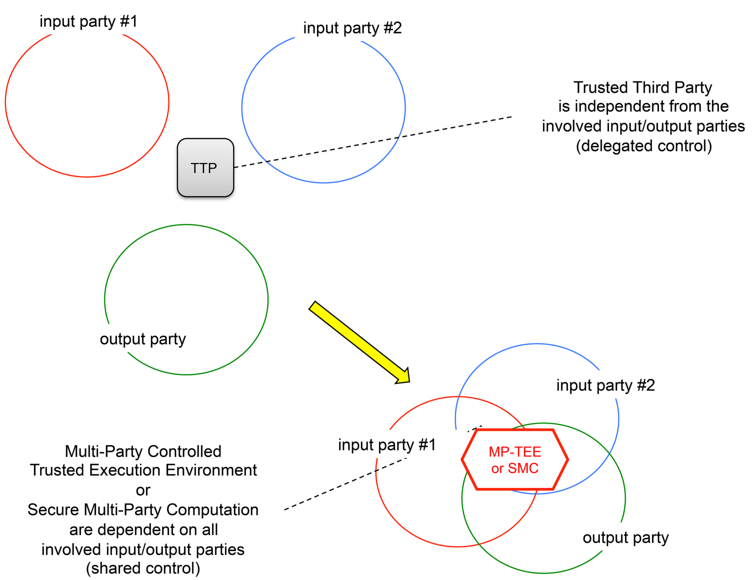


Fig. 3. Delegating control versus sharing control. The Trusted Third Party model (left) all parties must delegate control to an external entity. The technical solutions for Trusted Smart Statistics (like e.g. Multi-Party Controlled TEE or SMC) should instead aim to retain direct (non-exclusive) control among the key stakeholders

European Commission

- *Examples of applications*

Application domains

- B2G - Data from Private Data Holders (PDH)
  - **E.g. merge data from competing Mobile Network Operator (MNO)**
- C2G - Data from Citizens, Trusted Smart Surveys
  - **PET as tool for private computation (similar to Federated Learning concept)**
- G2G - Data from other public actors
  - **Different government agencies, administrative authorities from different countries, etc.**

European Commission

Example#1: Multi-MNO data integration

- *Input parties: the 3-5 Mobile Network Operators (MNO) in a same country - B2G*
- *Privacy of personal data + business sensitivity*
- *Output parties: Statistical Office & participating MNOs*

- *Computation goal: integrate data from individual MNO view without disclosing detailed data to competitors*
  - **Total counts of inbound roamers**(*)
  - **Join spatio-temporal distributions of mobile users across MNO**
  - …

(*) see e.g. https://sharemind.cyber.ee/mobile-phone-data-meets-sharemind-hi

Example#2: Trusted Smart Survey

- *Input parties: citizens participating voluntarily to the survey through their mobile devices (several 1000s) - C2G*
    - **passive sensor data and/or active replies to explicit queries**
- *Privacy of personal data (possibly very sensitive)*
- *Output parties: Statistical Office*

- *Goal: compute basic aggregate statistics*

See e.g. https://ec.europa.eu/eurostat/cros/content/trusted-smart-surveys-possible-application-privacy-enhancing-technologies-official-statistics-short-paper-sis-2020_en

Summary

- Privacy-Enhancing Technologies (PET) will have a role in the context of Trusted Smart Statistics (TSS)(*)

- PET as tools to improve **trust** by stakeholders (data providers, public) in how data are/will/can be used (for what, how, by whom).

- PET to deliver hard guarantees: make **technically unfeasible,** on top of **legally prohibited**, any deviation from agreed use.

- PET to enable transfer of the strictly necessary information, not whole data.

The GDPR sets out seven key principles:

- Lawfulness, fairness and transparency.
- **Purpose** limitation.
- Data minimisation.
- **Accuracy**.
- Storage limitation.
- **Integrity** and **confidentiality** (**security**)
- **Accountability**.

(*) *Trusted Smart Statistics: How new data will change official statistics*
Data & Policy journal, https://doi.org/10.1017/dap.2020.7

(**) Trusted smart statistics: Motivations and principles.
https://ec.europa.eu/eurostat/cros/system/files/sji190584.pdf

European Commission

Thanks for your attention

fabio.ricciato@ec.europa.eu