Quality for new data sources:

progress, challenges and directions of work

for the European Statistical System

Fabio Ricciato

European Commission, Eurostat
Unit A5 – Methodology; Innovation in Official Statistics

ISTAT Workshop on Methodologies for Official Statistics
Rome, 7 December 2023

Quality for new data sources: progress, challenges and directions of work for the European Statistical System

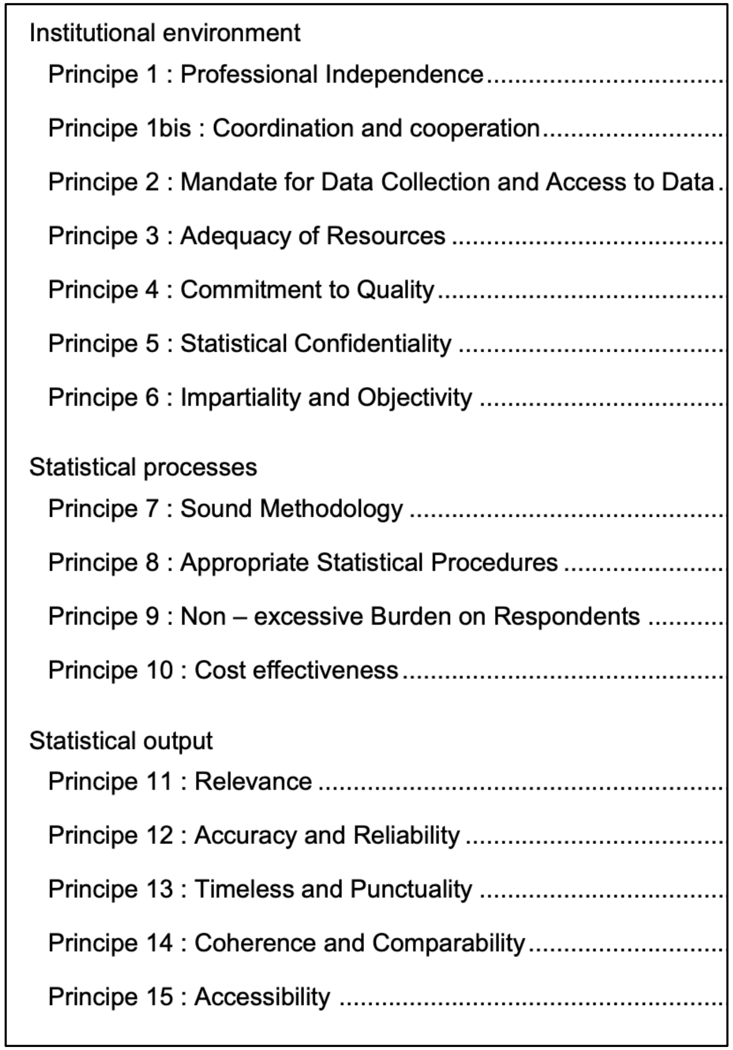
Caveat: The information and views set out in this presentation are those of the author and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Quality in OS - What?

Quality Assurance Framework of the European Statistical System

EUROPEAN
STATISTICAL
SYSTEM

Version 2.0



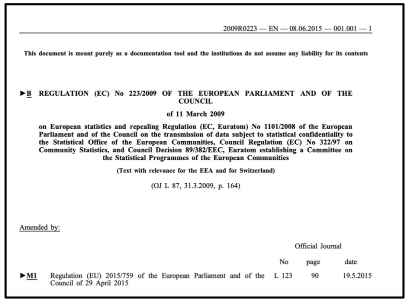
Institutional environment

Statistical processes

Statistical output

and

Quality in OS evolve …



2009R0223 — EN — 08.06.2015 — 001.001 — 1

This document is meant purely as a documentation tool and the institutions do not assume any liability for its contents

►B REGULATION (EC) No 223/2009 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 11 March 2009

on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities

(Text with relevance for the EEA and for Switzerland)

(OJ L 87, 31.3.2009, p. 164)

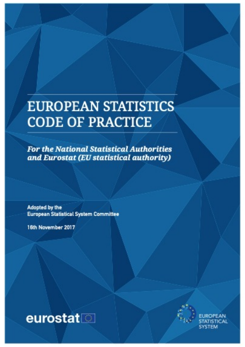
Amended by:

		Official Journal		
		No	page	date
►M1	Regulation (EU) 2015/759 of the European Parliament and of the Council of 29 April 2015	L 123	90	19.5.2015

Regulation 223/2009 on European Statistics

Proposal
Rev. 2023

Rev. 2015



EUROPEAN STATISTICS
CODE OF PRACTICE

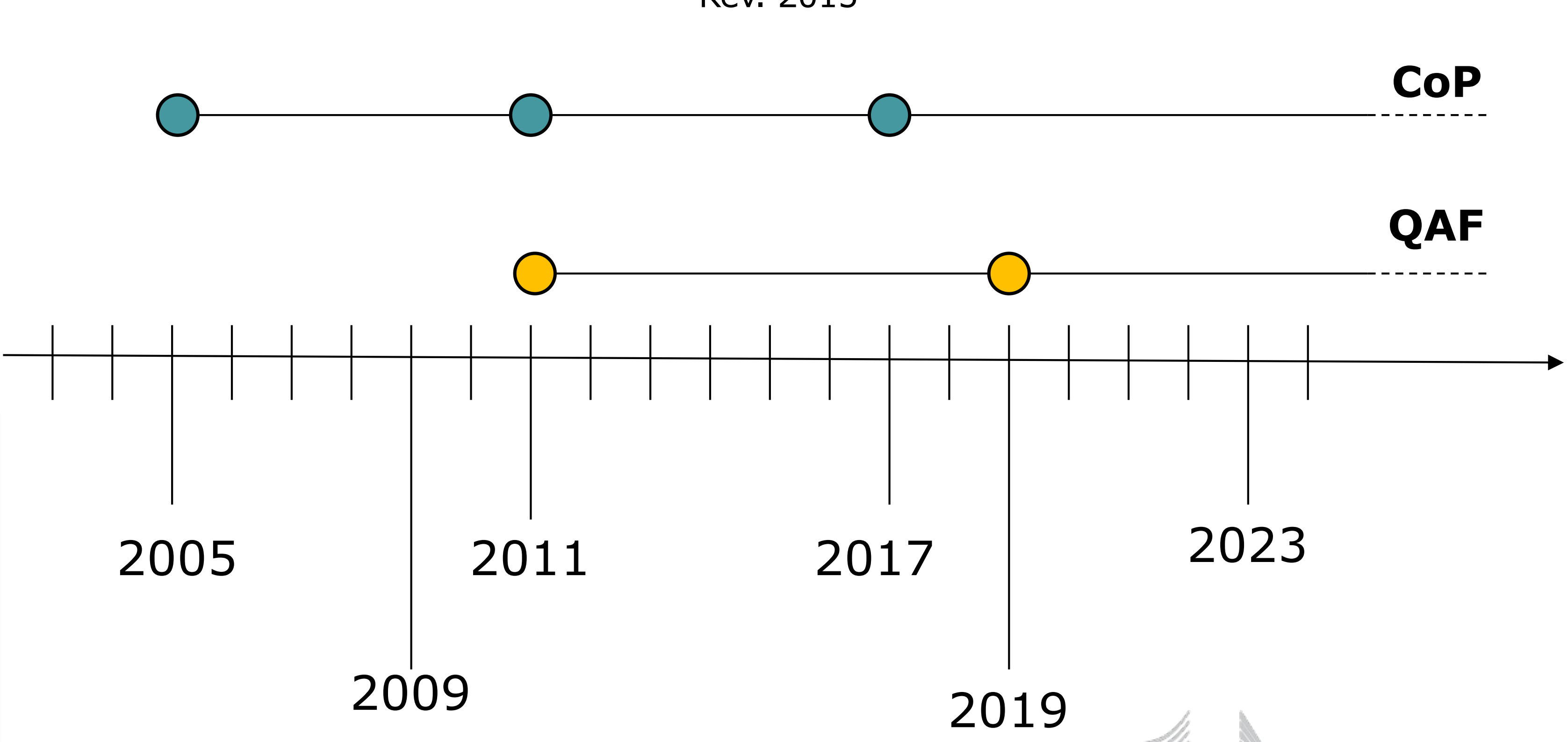
For the National Statistical Authorities and Eurostat (EU statistical authority)

Adopted by the
European Statistical System Committee

16th November 2017

eurostat

EUROPEAN
STATISTICAL
SYSTEM



Quality Assurance Framework of the
European Statistical System

EUROPEAN
STATISTICAL
SYSTEM

Version 2.0

European
Commission

Quality in OS - Why?

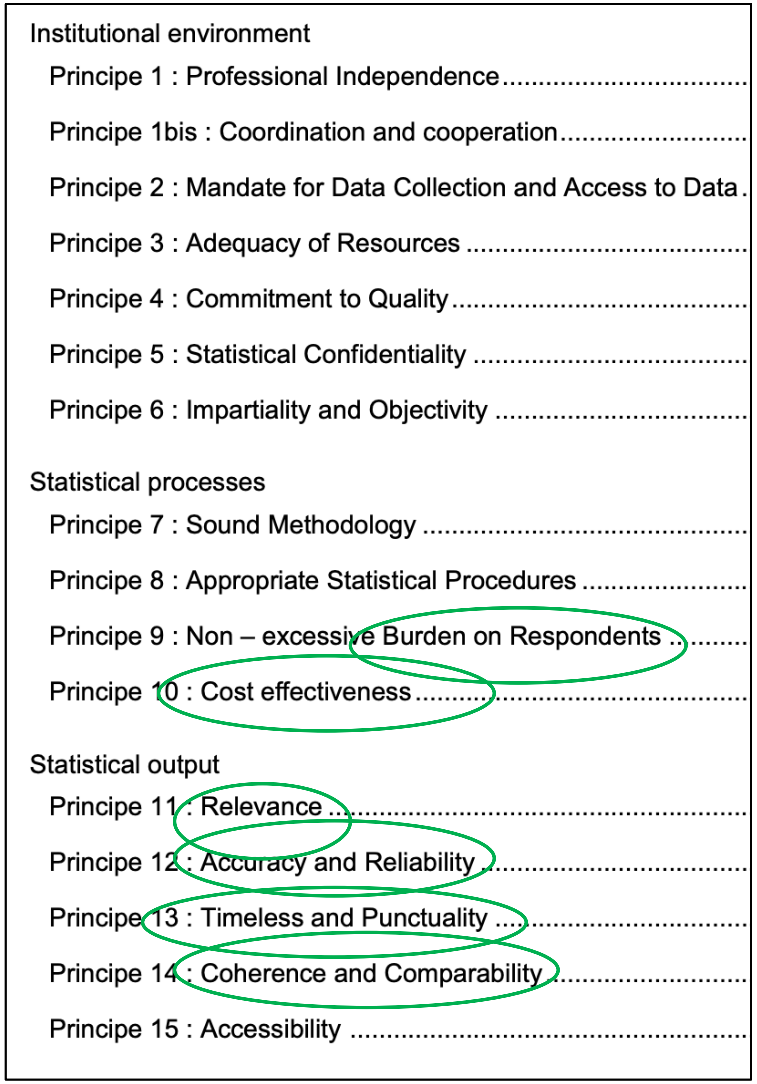
- *OS influence opinions, decisions …*
  - … by policymakers, citizens, businesses, researchers…

- ***Quality** as differentiator of OS vs. other stats*
  - OS ≠ commercial statistics, other public statistics
  - OS ≠ "experimental statistics"
    - One-off case study or short series, **no commitment** to regular production
    - Partial fulfillment of quality criteria, no full compliance

- ***O**fficial **S**tatistics = **Q**uality **S**tatistics*

QS

New data sources in OS: Opportunities…



Institutional environment

- Principe 1 : Professional Independence
- Principe 1bis : Coordination and cooperation
- Principe 2 : Mandate for Data Collection and Access to Data
- Principe 3 : Adequacy of Resources
- Principe 4 : Commitment to Quality
- Principe 5 : Statistical Confidentiality
- Principe 6 : Impartiality and Objectivity

Statistical processes

- Principe 7 : Sound Methodology
- Principe 8 : Appropriate Statistical Procedures
- Principe 9 : Non – excessive Burden on Respondents
- Principe 10 : Cost effectiveness

Statistical output

- Principe 11 : Relevance
- Principe 12 : Accuracy and Reliability
- Principe 13 : Timeless and Punctuality
- Principe 14 : Coherence and Comparability
- Principe 15 : Accessibility

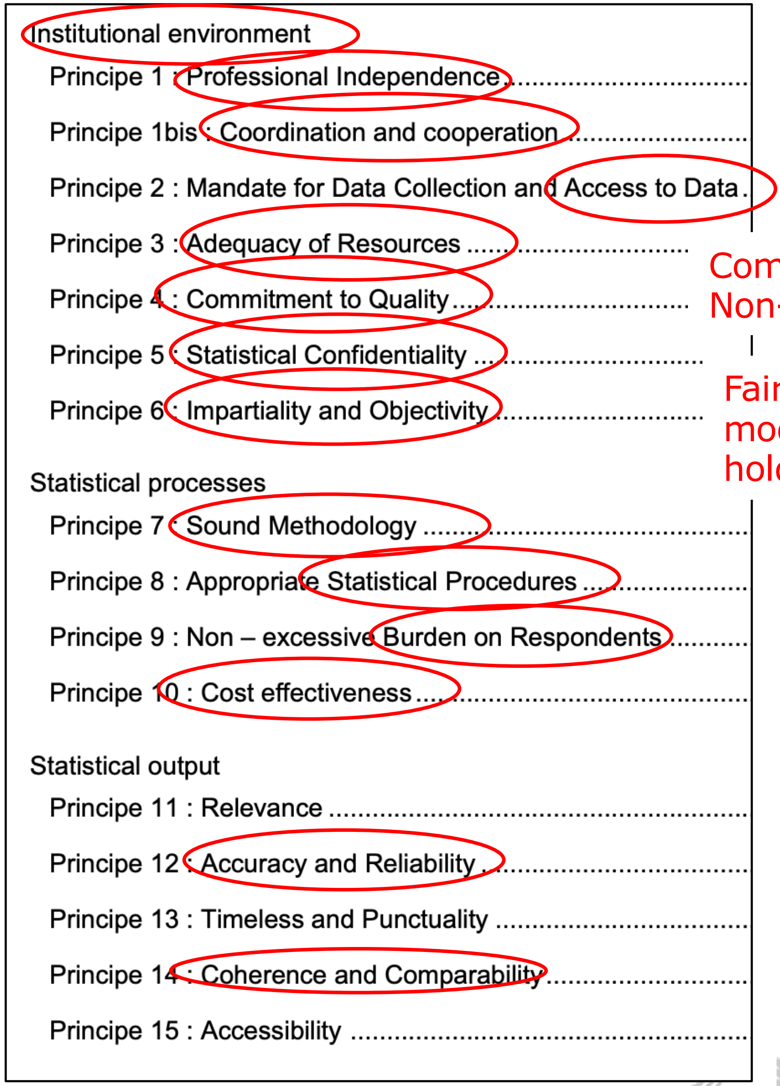
The potential gains

More, better, richer, timelier statistics *than what would be possible or feasible without new data sources*

augmenting, not replacing
OS based on traditional data

New data sources in OS: Challenges…

The actual pains



Institutional environment

Principe 1 : Professional Independence

Principe 1bis : Coordination and cooperation

Principe 2 : Mandate for Data Collection and Access to Data

Principe 3 : Adequacy of Resources

Principe 4 : Commitment to Quality

Principe 5 : Statistical Confidentiality

Principe 6 : Impartiality and Objectivity

Statistical processes

Principe 7 : Sound Methodology

Principe 8 : Appropriate Statistical Procedures

Principe 9 : Non – excessive Burden on Respondents

Principe 10 : Cost effectiveness

Statistical output

Principe 11 : Relevance

Principe 12 : Accuracy and Reliability

Principe 13 : Timeless and Punctuality

Principe 14 : Coherence and Comparability

Principe 15 : Accessibility

Public acceptance

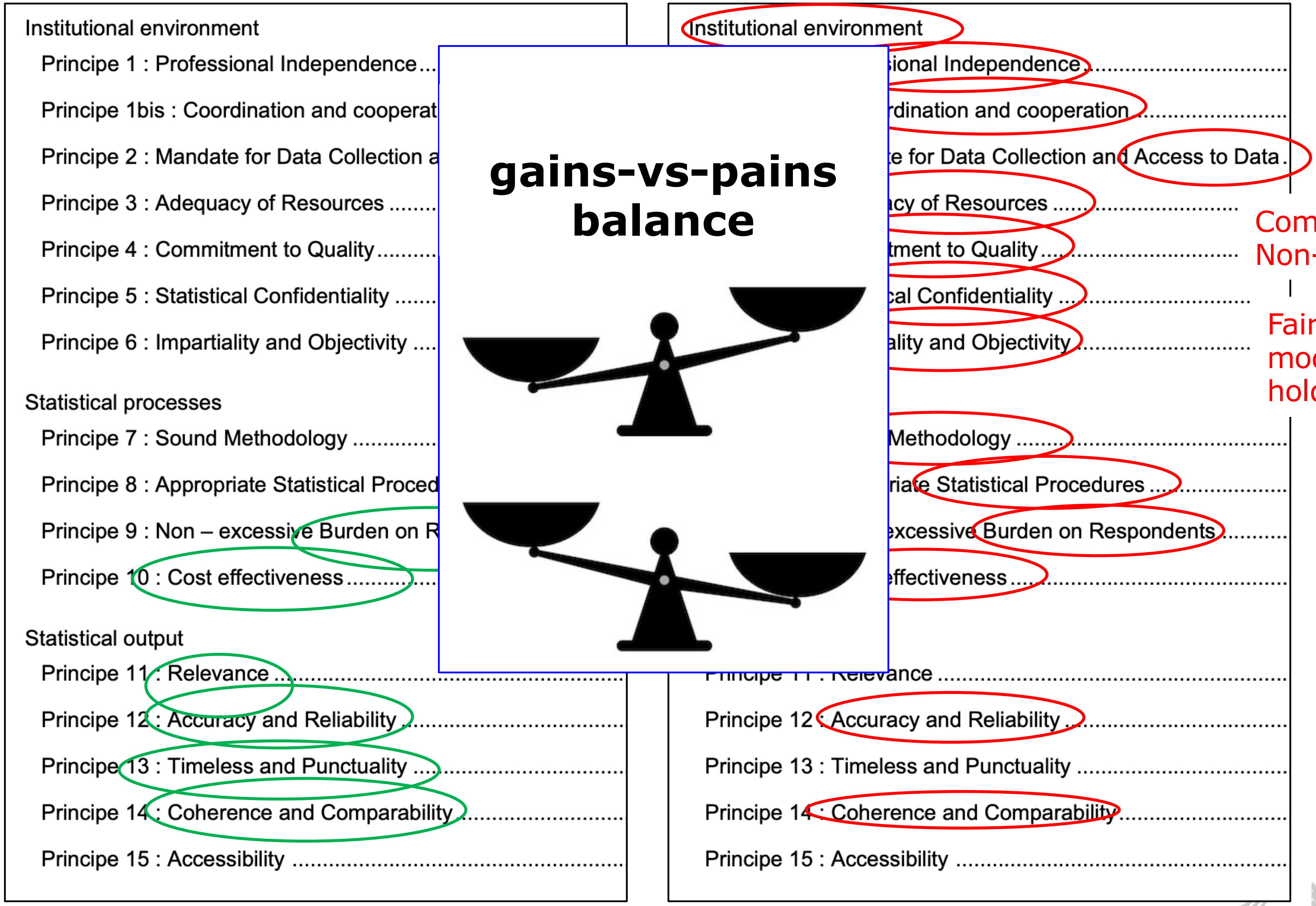
Compliance with
Non-statistical legislation

Fair partnership
models with data
holders

Reproducibility
of methods

Automatisation
Softwarisation

New data sources in OS – which ones?



Institutional environment

Principe 1 : Professional Independence...

Principe 1bis : Coordination and cooperat

Principe 2 : Mandate for Data Collection a

Principe 3 : Adequacy of Resources ....

Principe 4 : Commitment to Quality..........

Principe 5 : Statistical Confidentiality ....

Principe 6 : Impartiality and Objectivity ....

Statistical processes

Principe 7 : Sound Methodology ....

Principe 8 : Appropriate Statistical Proced

Principe 9 : Non – excessive Burden on R

Principe 10 : Cost effectiveness..............

Statistical output

Principe 11 : Relevance ....

Principe 12 : Accuracy and Reliability ....

Principe 13 : Timeless and Punctuality ....

Principe 14 : Coherence and Comparability..............................

Principe 15 : Accessibility ....

gains-vs-pains balance

Institutional environment

ional Independence..................................

rdination and cooperation ....

e for Data Collection and Access to Data.

cy of Resources ....

tment to Quality.......................................

cal Confidentiality ....

ality and Objectivity ....

Methodology ....

riate Statistical Procedures ....

xcessive Burden on Respondents..........

effectiveness...

Principe 11 : Relevance ....

Principe 12 : Accuracy and Reliability ....

Principe 13 : Timeless and Punctuality ....

Principe 14 : Coherence and Comparability...............................

Principe 15 : Accessibility ....

Public acceptance

Compliance with
Non-statistical legislation

Fair partnership
models with data
holders

Reproducibility
of methods

Automatisation
Softwarisation

Selection of data sources

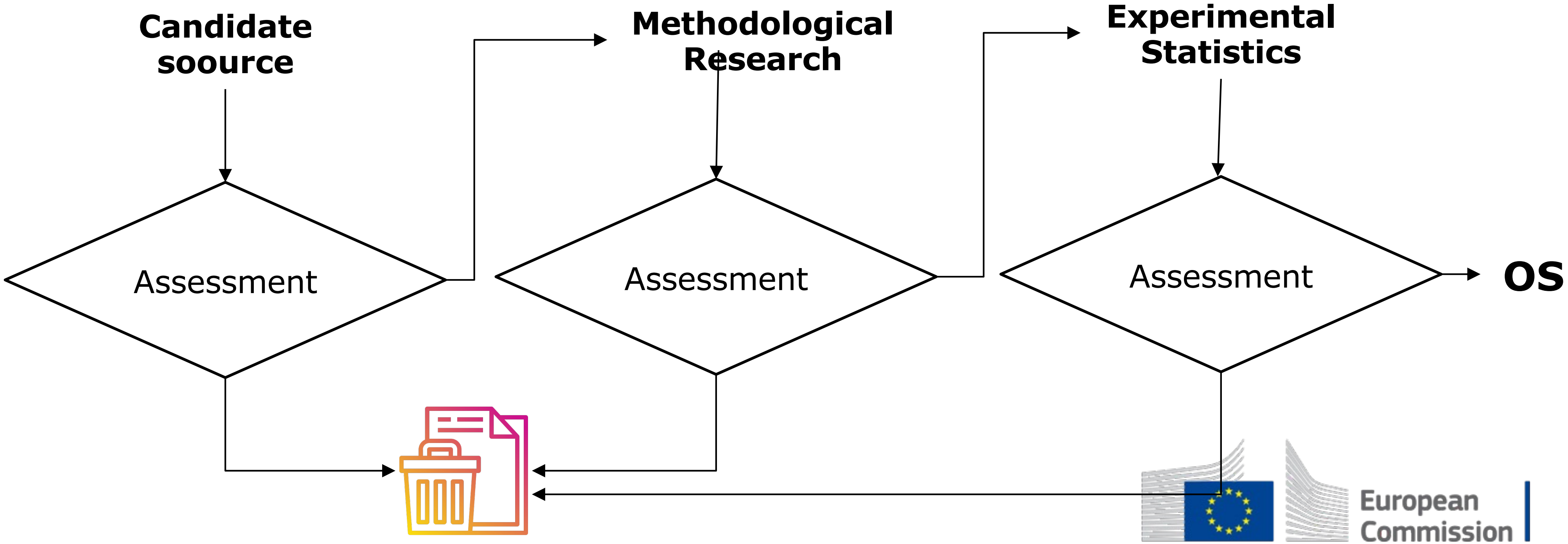
Be selective !

*New non-statistical data sources "**may** be reused for statistics" .. but also **may NOT!***

Decision based on gains-vs-pains balance …

… for some data sources the balance is negative …

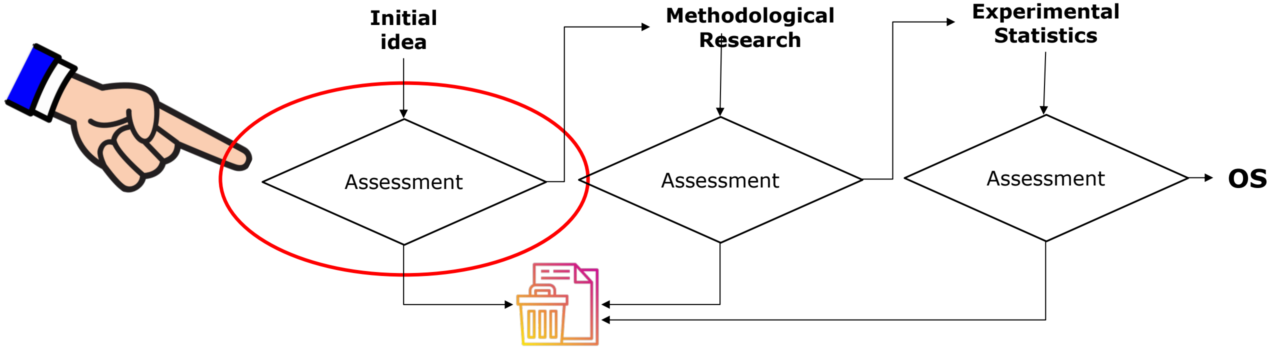
… or is negative today and could turn positive tomorrow …



Selection of data sources

Dimensions to be considered (among others)

- *Technological and/or market* ***penetration***
  - How many ”statistical units” can be reached?
  - What spatial coverage at European level?
- *Technological and/or market* ***stability***
  - Can we expect this data source to be there in 5-10-15 years?
  - How different will the data be in 5-10-15 years?
- ***Fragmentation/heteroegenity*** *of data formats/semantics*
  - Are there *de facto* or *de jure* standards?
  - Shall we deal with 1x, 10x, 100x or 1000x formats/APIs?



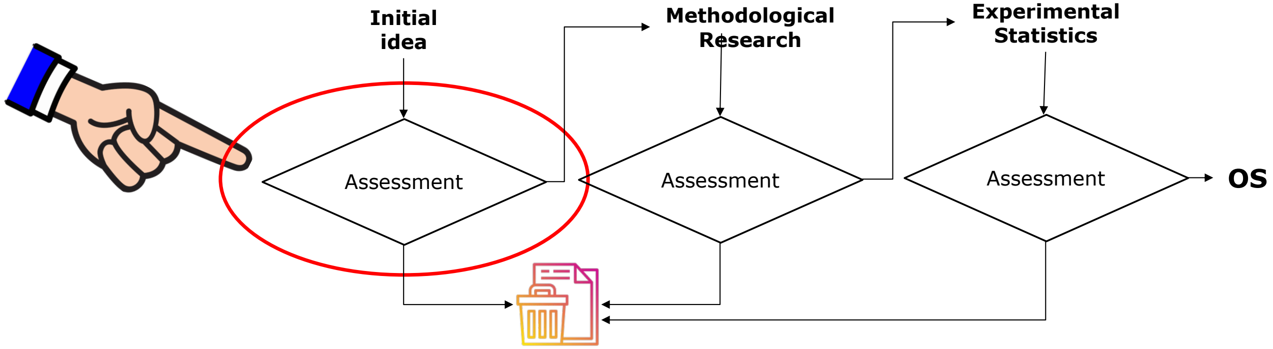
Selection of data sources

These dimensions depend on market and technology aspects, e.g.:

- *Market **structure** –vendors, integrators, adopters– who determines the data content and formats?*
- *Market **concentration** – one player, few players, many players?*

Do NOT depend on methodology or feedback from statistical users

→ it can be assessed at the beginning, even before research/experimental activities (landscaping analysis)



Classes of data sources

- *Methodologies and quality aspects cannot encompass whole set of "new data sources" but must be specific to "classes" of data sources*
- *ESSnet Big Data II – WP K* (*)

*"We believe that it is barely possible to write down meaningful **quality guidelines**, which can be applied to all kinds of new data sources. Instead, we group […] according to so called **data classes**, for which we write down data-class specific quality guidelines." (*)*

*[…] with new data sources, it becomes more important to do a data-class-wise quality assessment than to go through one general error type after the other. The reason is that **processes diverge hugely across data classes**, and so do **potential errors**.*

(*) ESSnet Big data II – WP K – deliverable K3 - 2020
https://cros-legacy.ec.europa.eu/content/wpk-milestones-and-deliverables_en

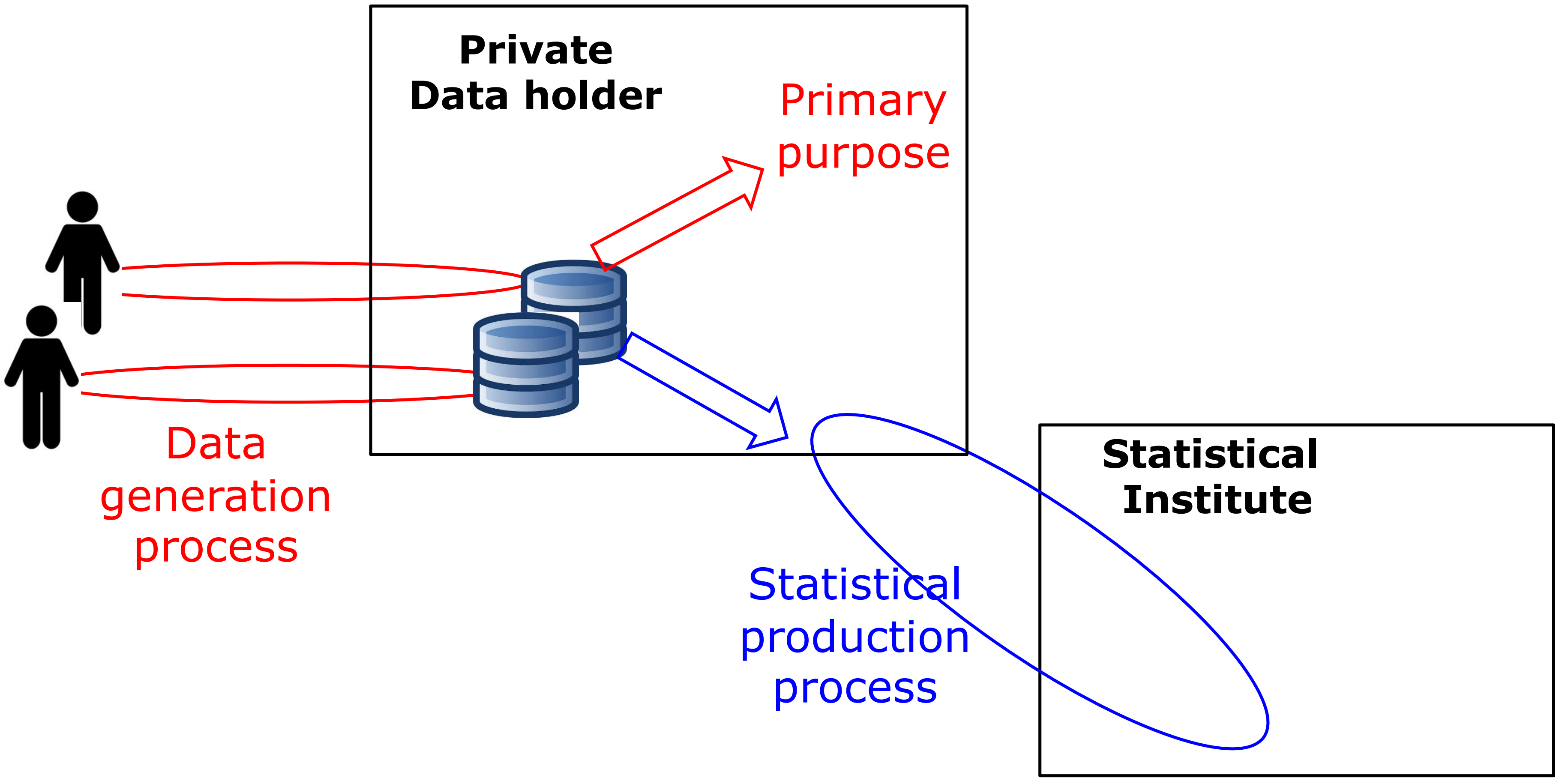
Focus on privately held data

Focus of this presentation on:

- *Digital traces, granular, sub-micro (nano-data)*
- *Generated primarily for non-statistical purposes*
- *Collected often by private companies (Privately Held Data)*
- *May be <u>re</u>used for statistics*

*Mutatis mutandis part of the following considerations apply also to (or can be inspirational for) other "**classes**" of new data sources (e.g. Earth Observation data, Internet data, …)*

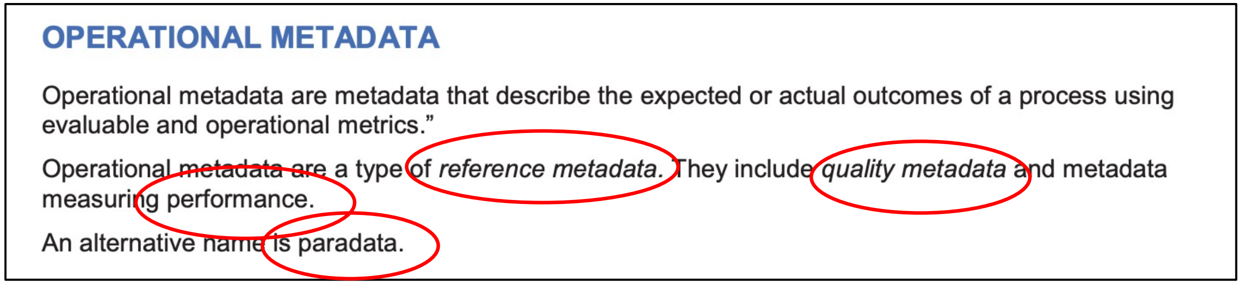
Reusing data held by somebody else is (also) a matter of processes …



… and of operational meta-data about the process and it's performance …

- *The data generation process is "dynamic"*
  - Changes in business, changes in technology → change in data
  - Customer churning, change in customer behaviour → change in data
  - Planned changes, e.g. system upgrade, new devices, new capabilities, new tariffs → change in data
  - Unplanned changes, e.g., system outage, errors, anomalies …
  - (NB: the above applies to the data generation, not to the data processing, and holds true regardless of how the data processing is split between PDH and NS)

All these "facts" and "events" affect input data and their quality, therefore must be identified and reported proactively (when possible) or at least detected and reported reactively



OPERATIONAL METADATA

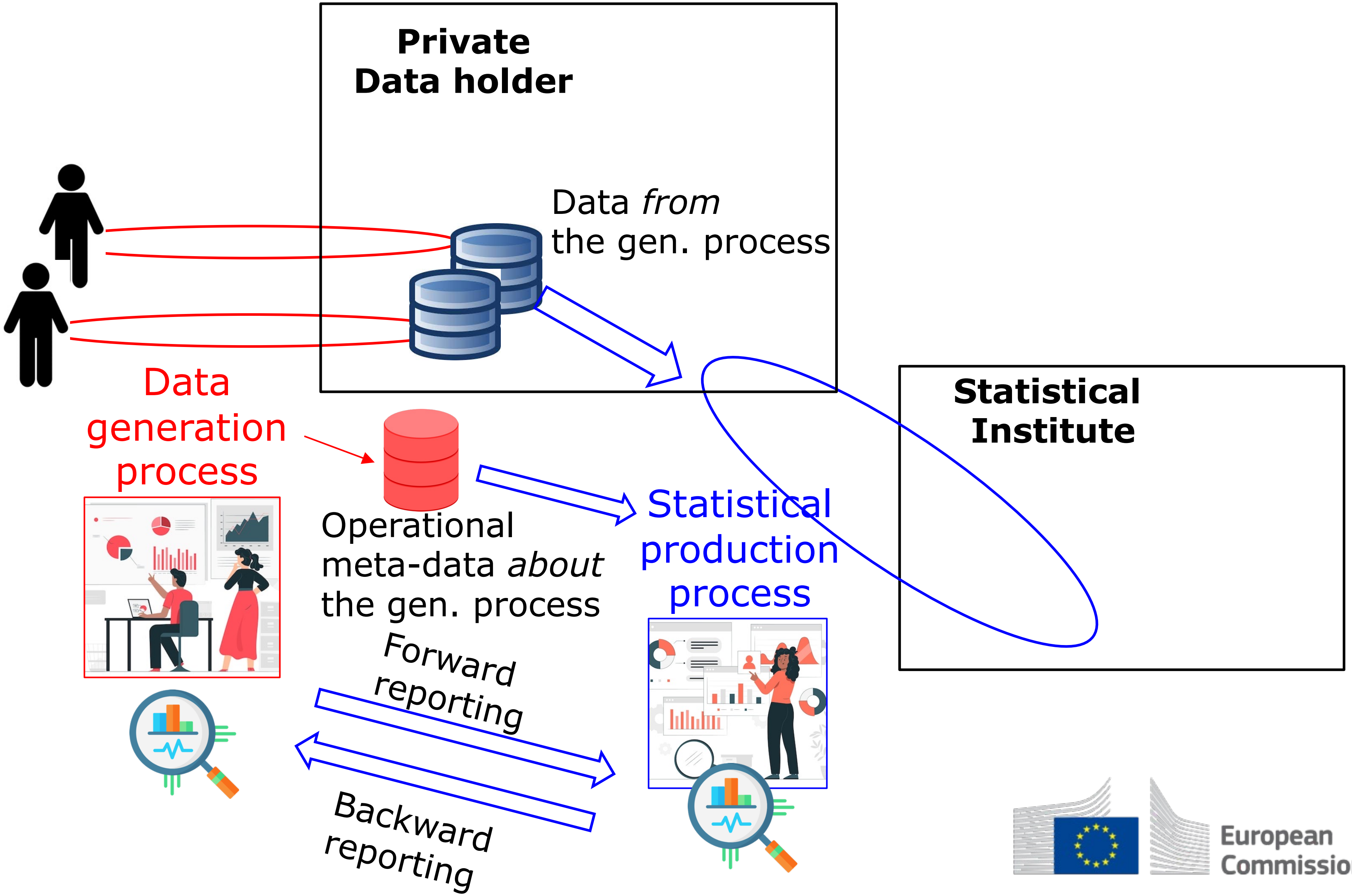
Operational metadata are metadata that describe the expected or actual outcomes of a process using evaluable and operational metrics."

Operational metadata are a type of *reference metadata.* They include *quality metadata* and metadata measuring performance.

An alternative name is paradata.

Source: ESS Handbook for quality and metadata reports
https://europa.eu/!3cWkFk

… and of operational meta-data about the process and it's performance …



Private Data holder

Data *from* the gen. process

Statistical Institute

Data generation process

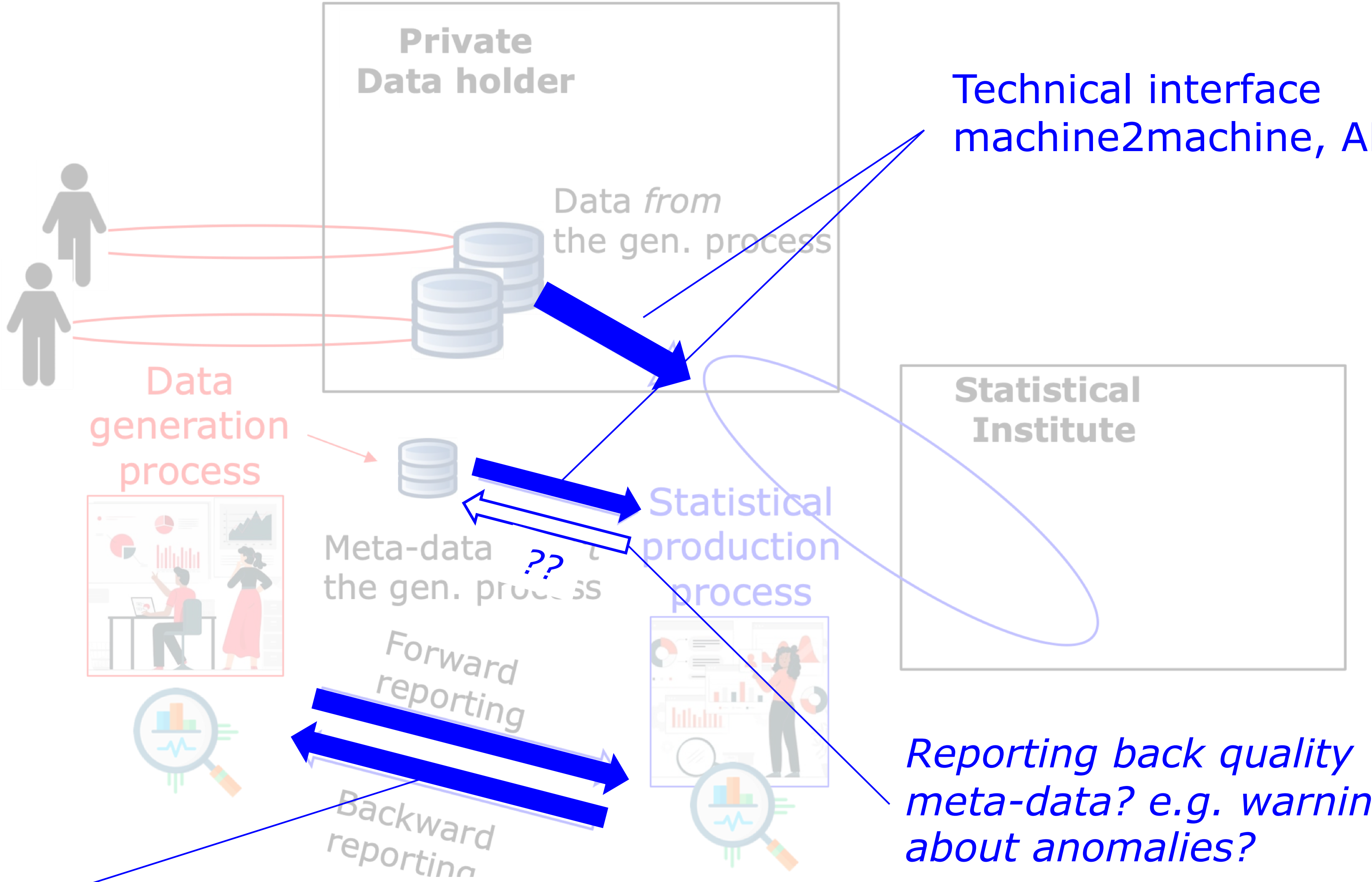
Operational meta-data *about* the gen. process

Statistical production process

Forward reporting

Backward reporting

… and interfaces, protocols, policies, etc. between NSI and data holders …



Private Data holder

Data *from* the gen. process

Data generation process

Meta-data

the gen. process

??

Statistical production process

Statistical Institute

Forward reporting

Backward reporting

Technical interface
machine2machine, API

Reporting back quality meta-data? e.g. warnings about anomalies?

Organisational interface – **bidirectional** (!)
Human2human: **forms** – **policies**

Communicating process meta-data

- *Communicate in both directions !!*
  - PDH-to-NSI: “Next week we plan a large system upgrade, will cause outages or generation of spurious data from time X to Y in region Z”
  - NSI-to-PDH “We detected an anomalous data pattern starting around time X that apparently affects region Z, can you please help us determine what is happening”?

- *Balance between under-reporting (bad for accuracy) and over-reporting (bad for burden)*

How to motivate the data holders?

- ***Incentives?***
- ***Legislative obligations?***
- ***Cost compensation?***

- *All of the above*
  - according to the Expert Group on facilitating the use of new data sources for official statistics (2022) https://europa.eu/!JGR3Gx

Empowering society by reusing privately held data for official statistics — A European approach

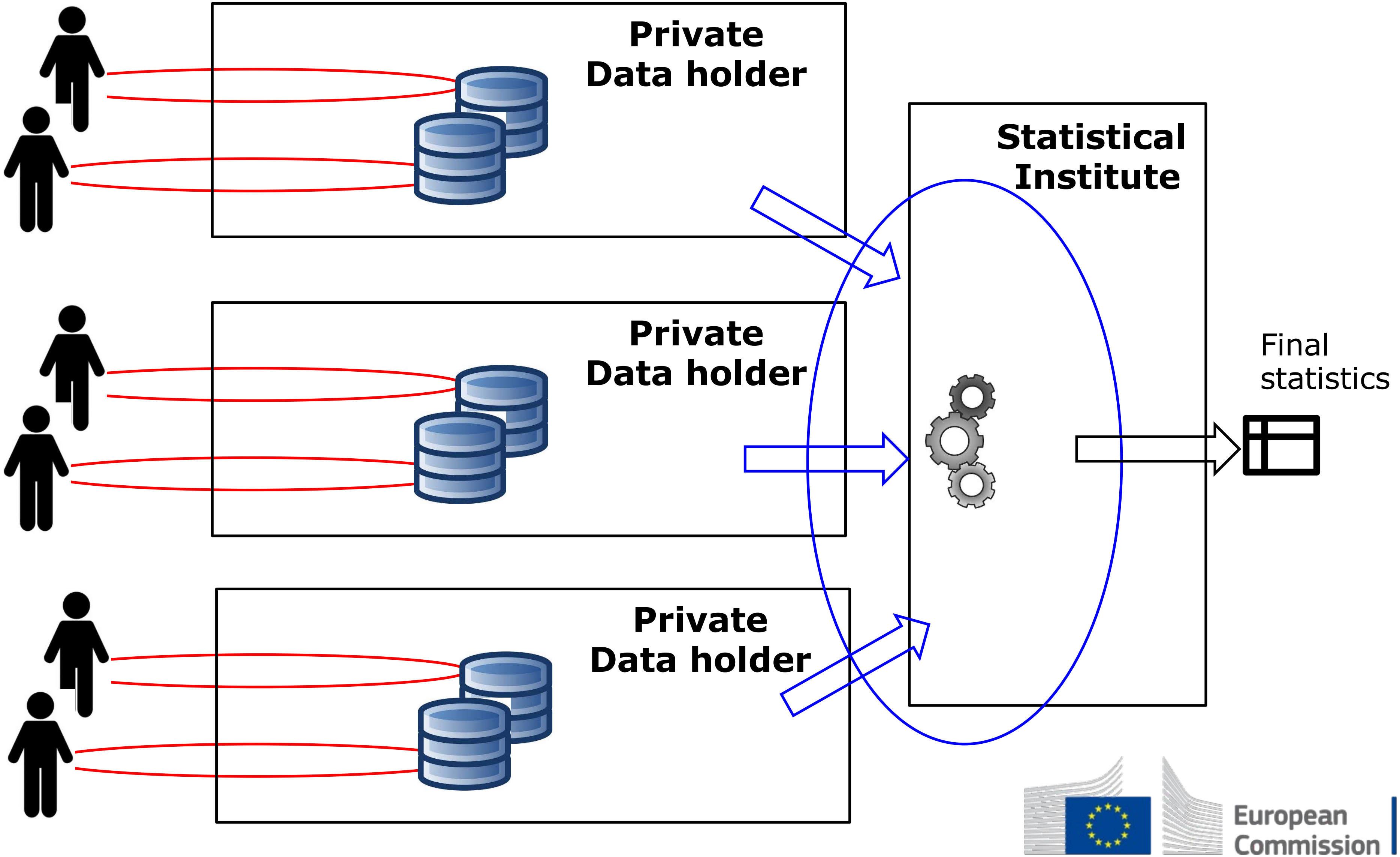
FINAL REPORT PREPARED BY THE HIGH-LEVEL EXPERT GROUP ON FACILITATING THE USE OF NEW DATA SOURCES FOR OFFICIAL STATISTICS

2022 edition

STATISTICAL REPORTS

eurostat

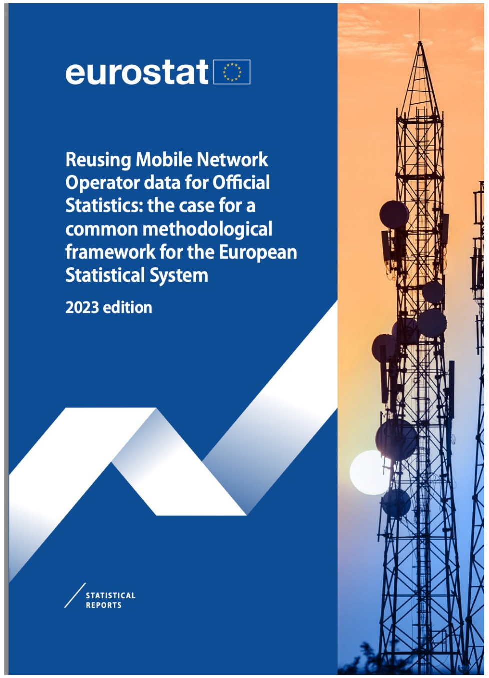
Integration of data from multiple data holders improves quality



European Commission

Integration of data from multiple data holders improves quality

- *Better representativeness of the total population, mitigate effects of population coverage bias in the final statistics*
    - Privately Held Data refer to "customers", not citizens
- *Improved temporal stability, mitigation of customer churning*
- *Mitigate sensitivity to provider-specific aspects of data generation*
- *Improved robustness to anomalies, outages, partial or complete disruptions of data provision*
- *Equal treatment of all data providers in the same business sector ("level playing field")*
- *Easier protection of business-sensitive information from individual data providers in the final statistics*



eurostat

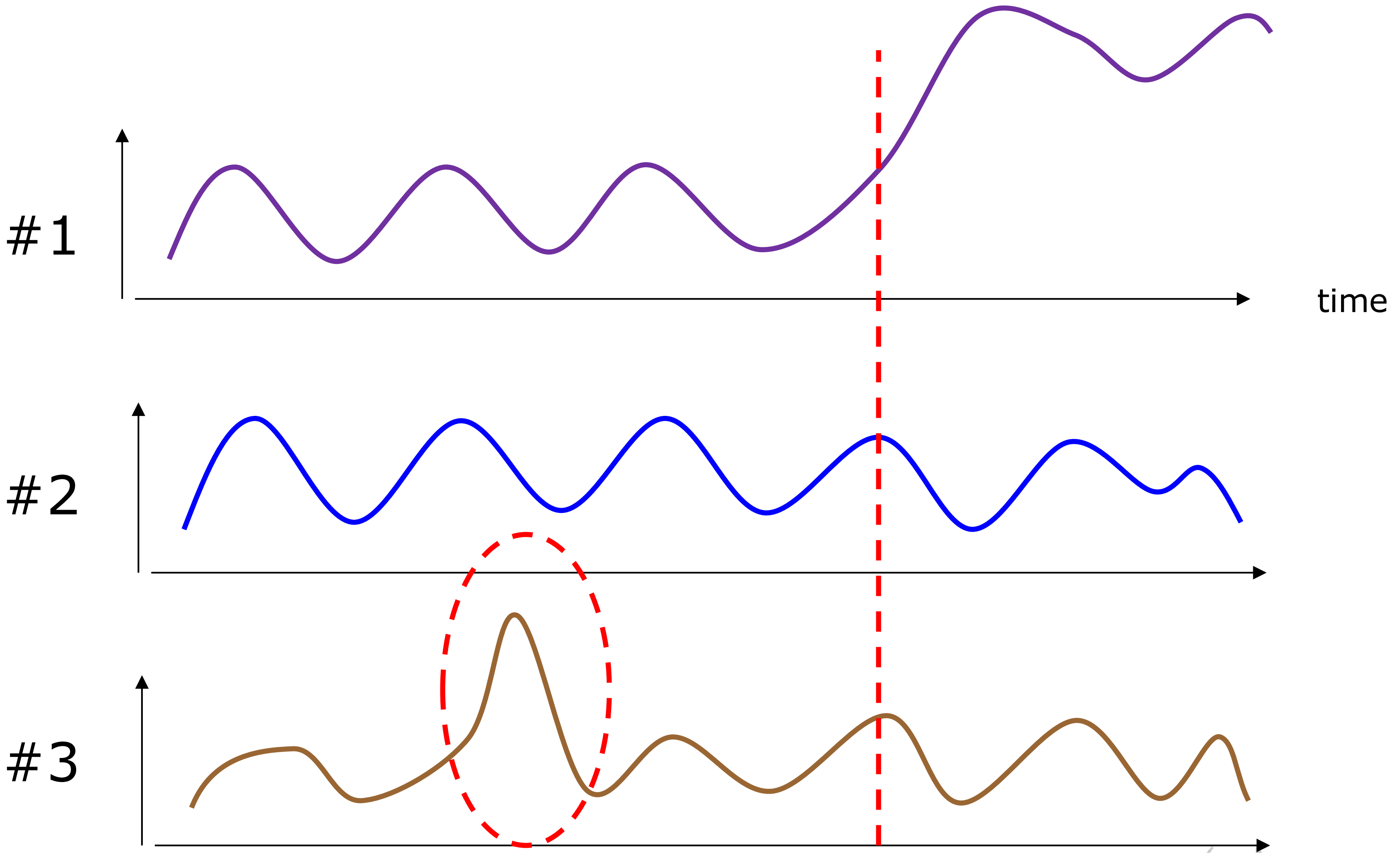
Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System

2023 edition

STATISTICAL REPORTS

Source: Position paper by ESS Task Force on MNO data "*Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System*", *2023* https://europa.eu/!KbdVG4

Consistency and plausibility checks across data providers



European Commission

Integration of non-statistical "big" data and statistical data

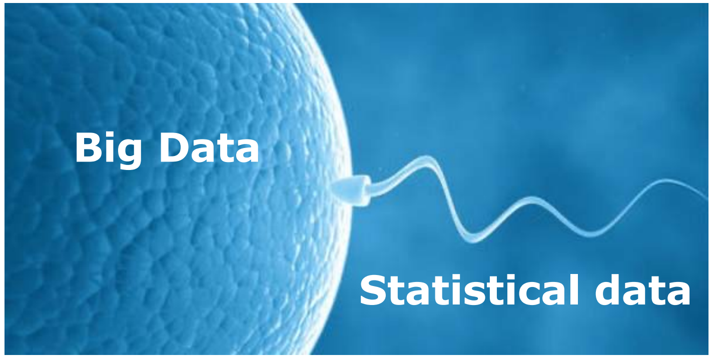
- *Data produced for business purposes refer to "customers" and oftent to "customers' devices"*
  - Observed population does not map 1:1 to target population
  - Observed population is not a representative sample of whole population
  - Coverage gaps, multiple counting, coverage bias
- *Mapping not static*
  - customer churning
  - change in user behaviour
- *Statistics based solely on a single "big data" source maybe inaccurate and unstable*
- *Big Data may lack variables of interest for statistics*

Integration of non-statistical "big" data and statistical data

- *Combination of non-statistical "big" data sources with (small?) statistical data may deliver the best of both*

  - From ”big data”: timeliness (near real-time), spatio/temporal detail ("interpolation"), temporal continuity and spatial coverage, variables derived from "objective" observations

  - From statistical data: correct projection to target population (mitigation of bias, multiple counting, coverage gaps), additional variables of interest not observed by big data



Fertilise big data by statistical data

Augment statistical data with big data

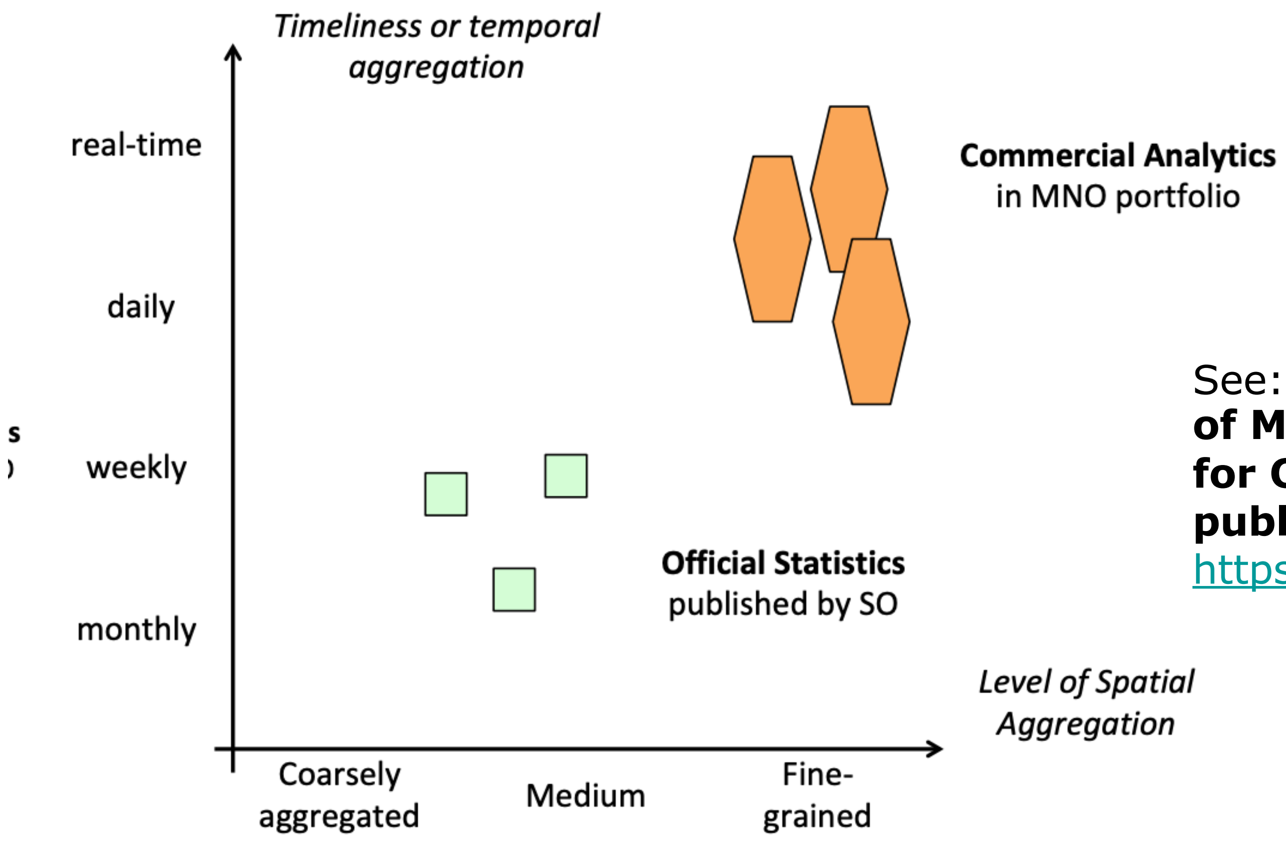
Integration of non-statistical "big" data and statistical data

- *More balanced positioning of NSI and private data holders*
  - Not provider-consumer, but win-win partnerhship

- *No derogations to statistical confidentiality!*
  - Statistical data cannot be used for non-statistical purposes
  - Statistics integrating data from **multiple providers *and* statistical data** produces an authoritative final statistics that may serve as "calibration reference" also for individual providers
  - Intermediate aggregate non-personal data may be shareable back with individual providers
  - Privacy-Enhancing Technologies may help to protect data confidentiality

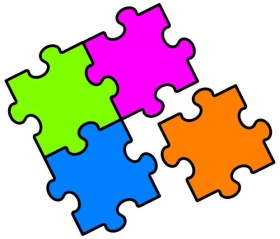
- *Reassert the role of NSIs and statistical surveys in the new data-rich ecosystem*
  - Survey still needed, but smaller and less burdensome if combined with big data, for better/richer/timelier final statistics

Commercial Analytics may coexist with, and even get reinforced by Official Statistics

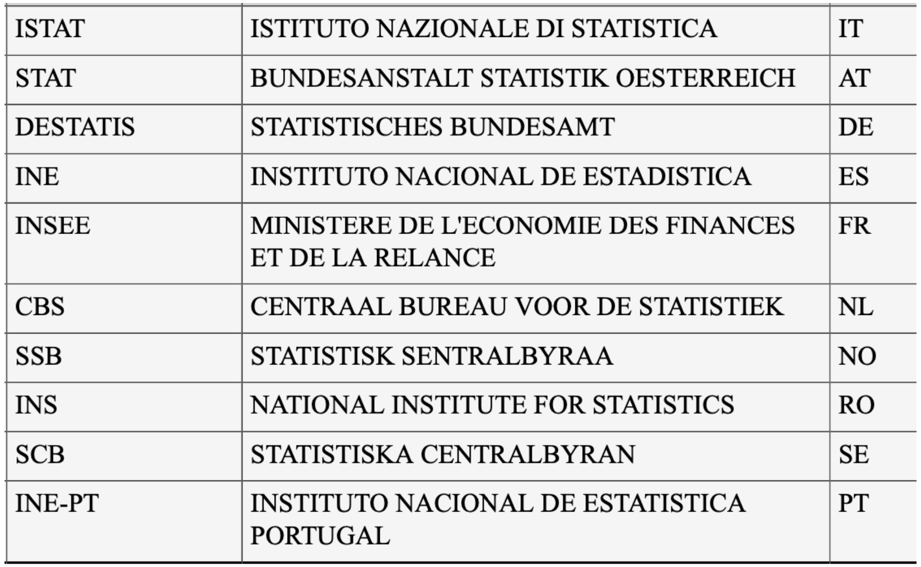

Timeliness or temporal aggregation
real-time
daily
weekly
monthly
s
)
Commercial Analytics in MNO portfolio
Official Statistics published by SO
Level of Spatial Aggregation
Coarsely aggregated
Medium
Fine-grained

See: DGINS 2018 paper **Processing of Mobile Network Operator data for Official Statistics: the case for public-private partnerships** https://zenodo.org/records/10246468

Integration of MNO and non-MNO data



- *ESSnet METH-TOO: **research grant** focusing on the combination of MNO and non-MNO data*
  - Started officially on 1st November 2023 for 2 years, 900 keur
  - Consortium of 10x NSI, ISTAT coordinator
- *Project Objectives*
  - *WP1 - Landscaping analysis of candidate non-MNO data sources lead FR*
  - *WP2 - Developing formal methods and open-source tools for integration of MNO and non-MNO data lead NO*
  - *WP3- Proof of concept of ad-hoc survey to improve MNO data lead AT*



ISTAT	ISTITUTO NAZIONALE DI STATISTICA	IT
STAT	BUNDESANSTALT STATISTIK OESTERREICH	AT
DESTATIS	STATISTISCHES BUNDESAMT	DE
INE	INSTITUTO NACIONAL DE ESTADISTICA	ES
INSEE	MINISTERE DE L'ECONOMIE DES FINANCES ET DE LA RELANCE	FR
CBS	CENTRAAL BUREAU VOOR DE STATISTIEK	NL
SSB	STATISTISK SENTRALBYRAA	NO
INS	NATIONAL INSTITUTE FOR STATISTICS	RO
SCB	STATISTISKA CENTRALBYRAN	SE
INE-PT	INSTITUTO NACIONAL DE ESTATISTICA PORTUGAL	PT

Softwarization of statistical methods

- ***Volume*** *and* ***complexity*** *of granular data*

- ***Automation*** *of data processing*
  - Human work shifts from ‘executing instructions’ to ‘formulating instructions’ to be executed by machines

- *Data processing to be represented in* ***formal languages***
  - Programming languages, schema, ontologies … code!

Automatization → ***Softwarization*** *of statistical methodologies**

(*) A reflection on methodological sensitivity, quality and transparency in the processing of new ‘big’ data sources, Q2022 conference, Vilnius
https://zenodo.org/records/10246446

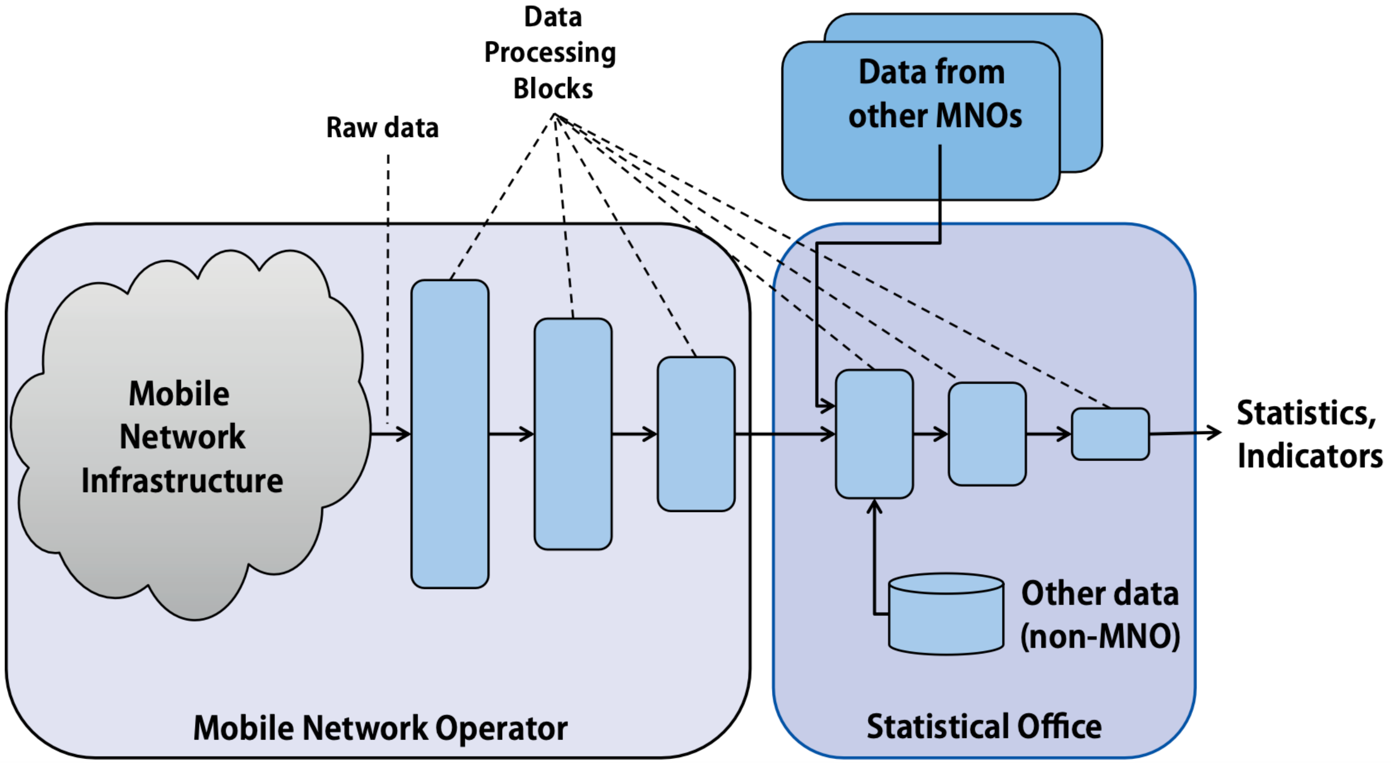
Softwarization of statistical methods

- *Opportunities and implications*
  - Source code as reference metadata !
  - Open-source code release to abide to methodological transparency
  - Reproducibility of methods (≠ replicability of results)
  - Independent auditability of methods, collaborative improvement
  - Methodological development may learn from (and import) established practices in complex software development (e.g. modularity, collaborative development, versioning)
  - Quality of (reference) software as natural component of methodological quality
  - Sharing code with other NSIs and/or data providers, ease harmonisation, pool resources

See also “Standardisation of methods and processes” presentation given at ISTAT Workshop on Methodologies for Official Statistics, Dec’22
https://www.istat.it/it/files/2022/11/4_3_Slide_Ricciato.pdf

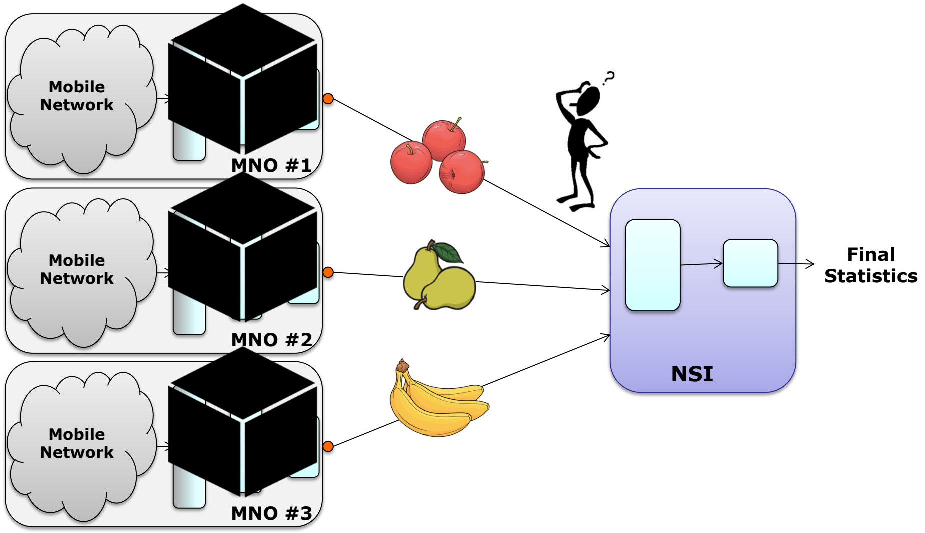
Standardisation of end-to-end data processing workflows

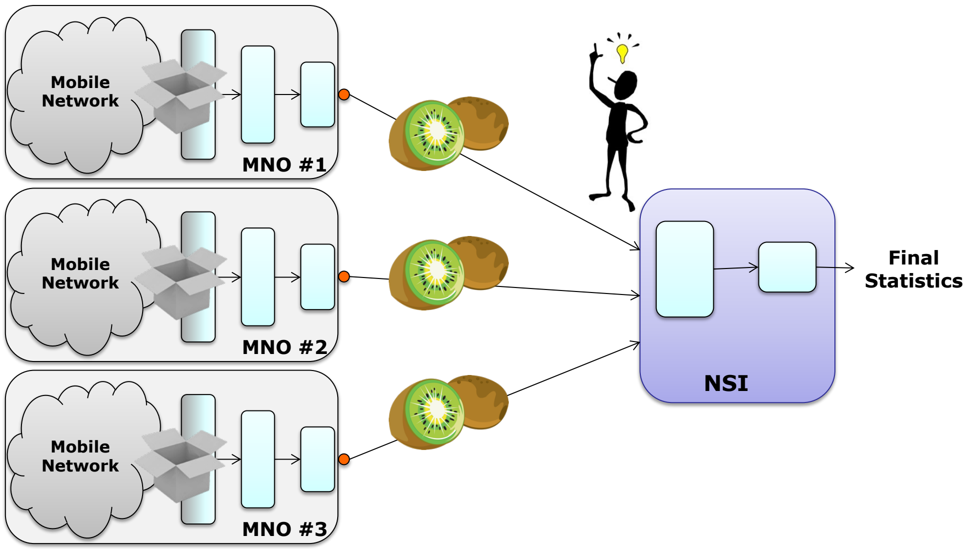
- *Detailed and non-ambiguous representation of operations and data structures in formal language …*
- *… (co)defined by NSI even if they are executed at data provider's premises*



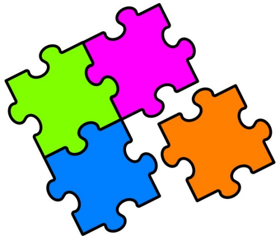
Softwarisation & standardisation

- *Common/standard methodologies and definitions necessary to achieve **comparable** and **combinable** results*
- *Softwarisation and Standardisation reinforce each other*





Multi-MNO project

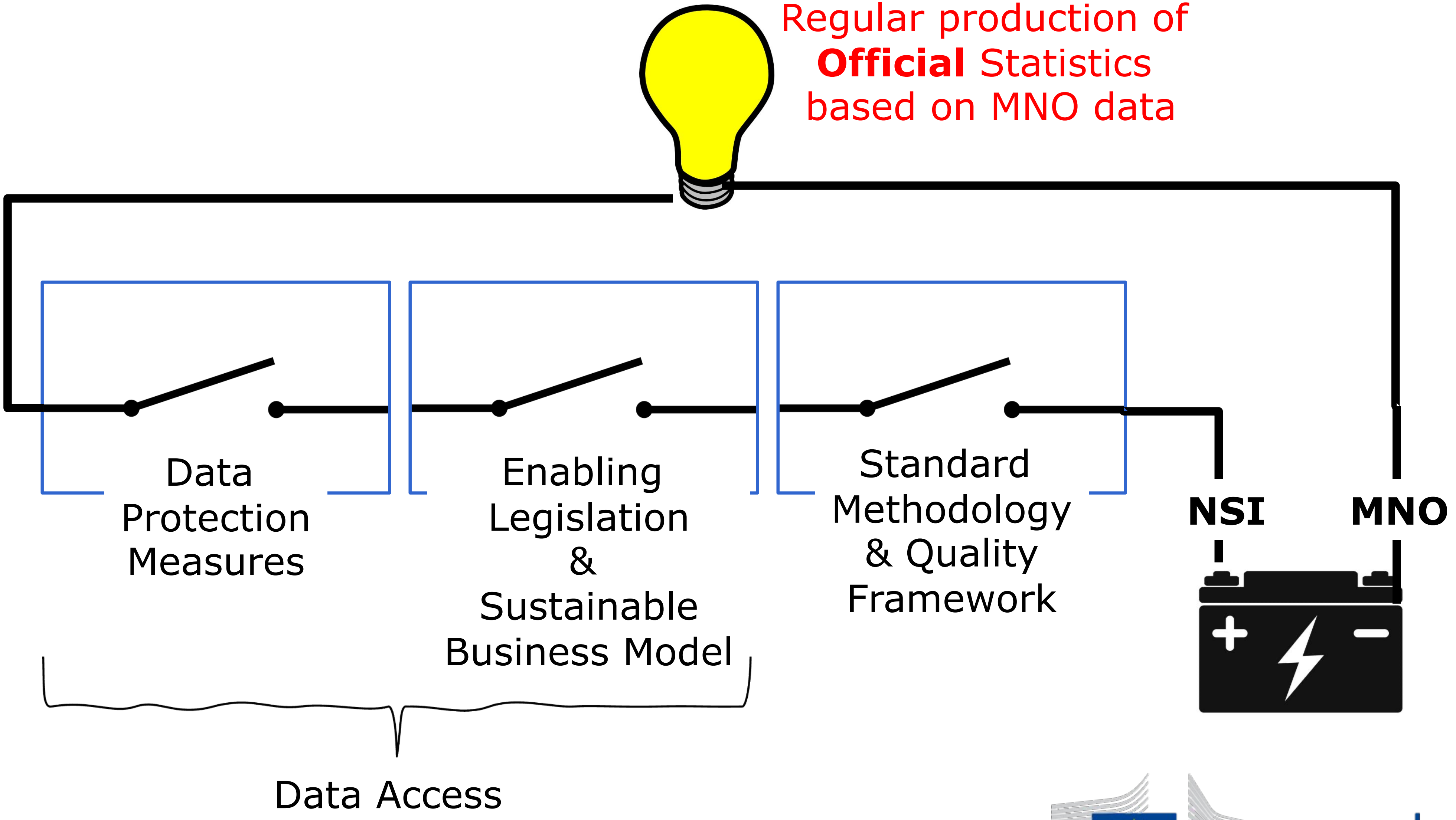


- ***Co-development partnership***
  involving NSI experts and industry experts working together
  - Started in Jan’23 for 2.5 years, until mid 2025
  - Public procurement procedure based on open call for tenders, 1.2 Mio
- *Project Objectives*
  - *Develop a first proposal for an open end-to-end **methodological** framework and associated **quality framework and guidelines**, with focus on an initial selection of of use-cases*
  - *Open-source reference software pipeline implementing the proposed methodological framework;*
  - *Practical demonstration of the processing pipeline on real-world data across 5x MNOs in 4x EU countries*
- *Consortium*
  - *GOPA (Germany, consortium leader)*
  - *2x Industry partners: NOMMON (Spain), POSITIUM (Estonia)*
  - *2x NSI: CBS (Netherlands), ISTAT (Italy)*
  - *5X MNOs: Orange Spain, Vodafone Spain, Vodafone Italy, A1 Slovenia, POST Lux.*

The vision

- *In 202x MNO data are (re)used for regular production of **official statistics***
  - Not merely “experimental statistics”…
  - Data from multiple MNOs in each country and across countries - Multi-MNO
  - Combined with statistical data
  - Processed according to **standard methodologies** and **transparent quality criteria** defined at EU level, by the ESS in collaboration with industry
  - Evolvable methodological framework
  - Processed (at least partly) at MNOs premises
  - Built-in privacy protection measures defined at EU level in consultation with EDPS/EDPB

Series of challenges need to be worked out in parallel



Regular production of **Official** Statistics based on MNO data

Data Protection Measures

Enabling Legislation & Sustainable Business Model

Standard Methodology & Quality Framework

Data Access

Composing the puzzle

ESSnet Big Data II – WPI & WPK

eurostat

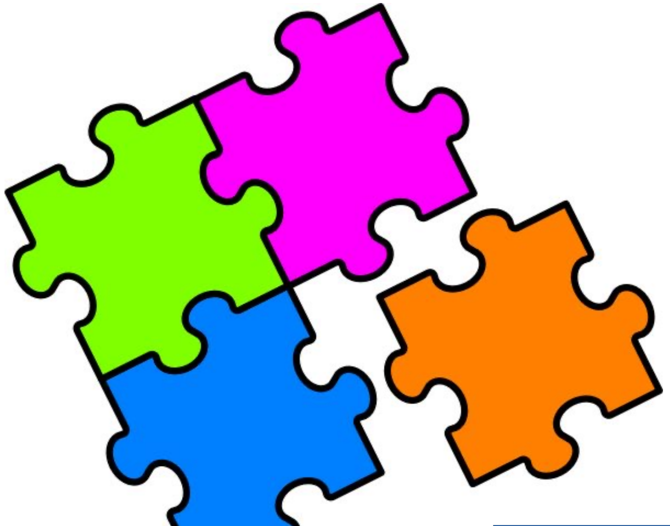
Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System

2023 edition

STATISTICAL REPORTS

TF MNO

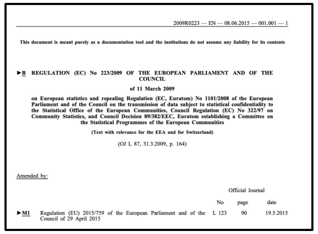
Multi-MNO project



Other projects on
Privacy-Enhancing
Technologies

Reserch Grant
ESSnet METH-TOO

Rev. 223



2009R0223 — EN — 08.06.2015 — 001.001 — 1

This document is meant purely as a documentation tool and the institutions do not assume any liability for its contents

►B REGULATION (EC) No 223/2009 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 11 March 2009

on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities

(Text with relevance for the EEA and for Switzerland)

(OJ L 87, 31.3.2009, p. 164)

Amended by:

		Official Journal		
		No	page	date
►M1	Regulation (EU) 2015/759 of the European Parliament and of the Council of 29 April 2015	L 123	90	19.5.2015

Empowering society by reusing privately held data for official statistics — A European approach

FINAL REPORT PREPARED BY THE HIGH-LEVEL EXPERT GROUP ON FACILITATING THE USE OF NEW DATA SOURCES FOR OFFICIAL STATISTICS

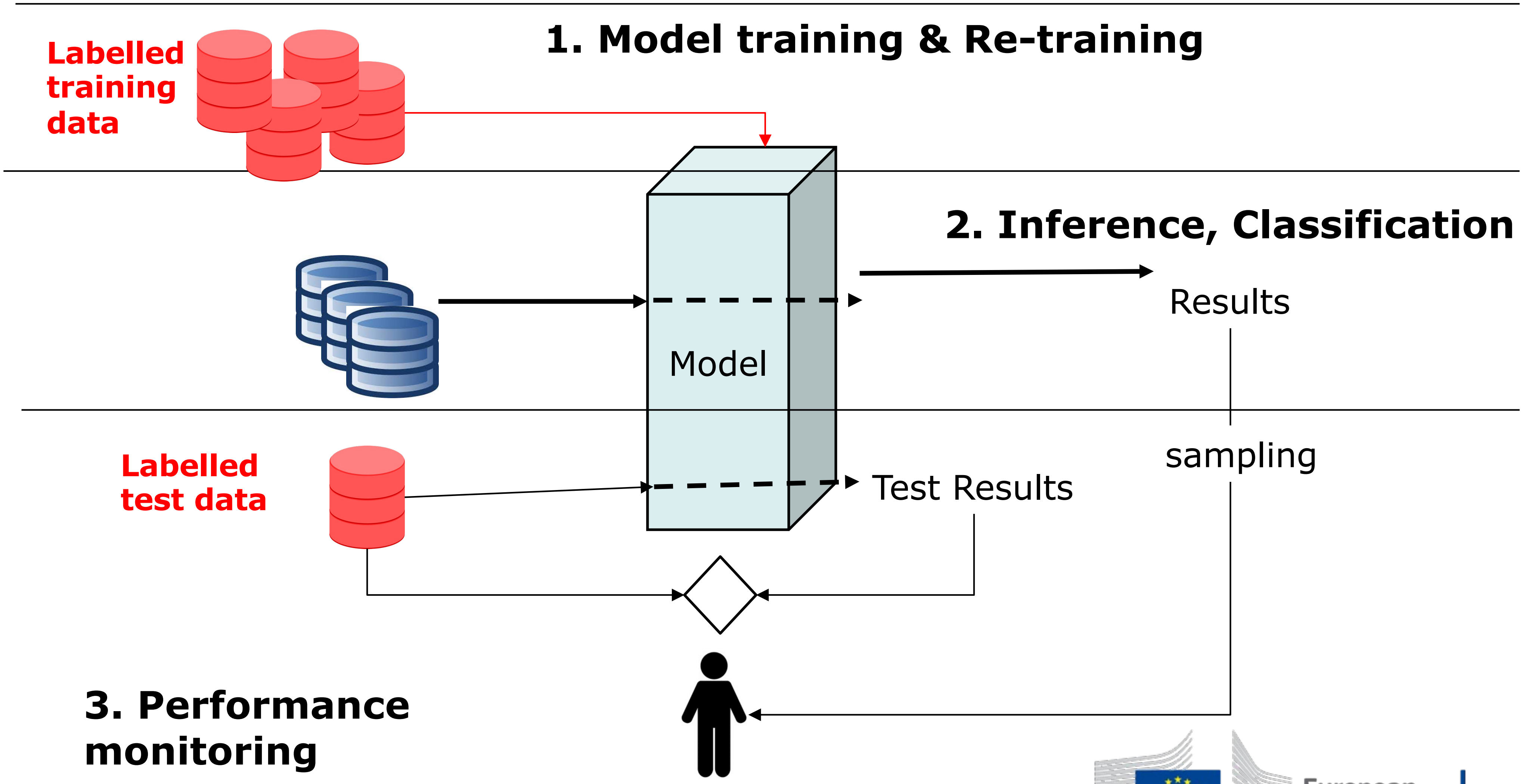
2022 edition

STATISTICAL REPORTS

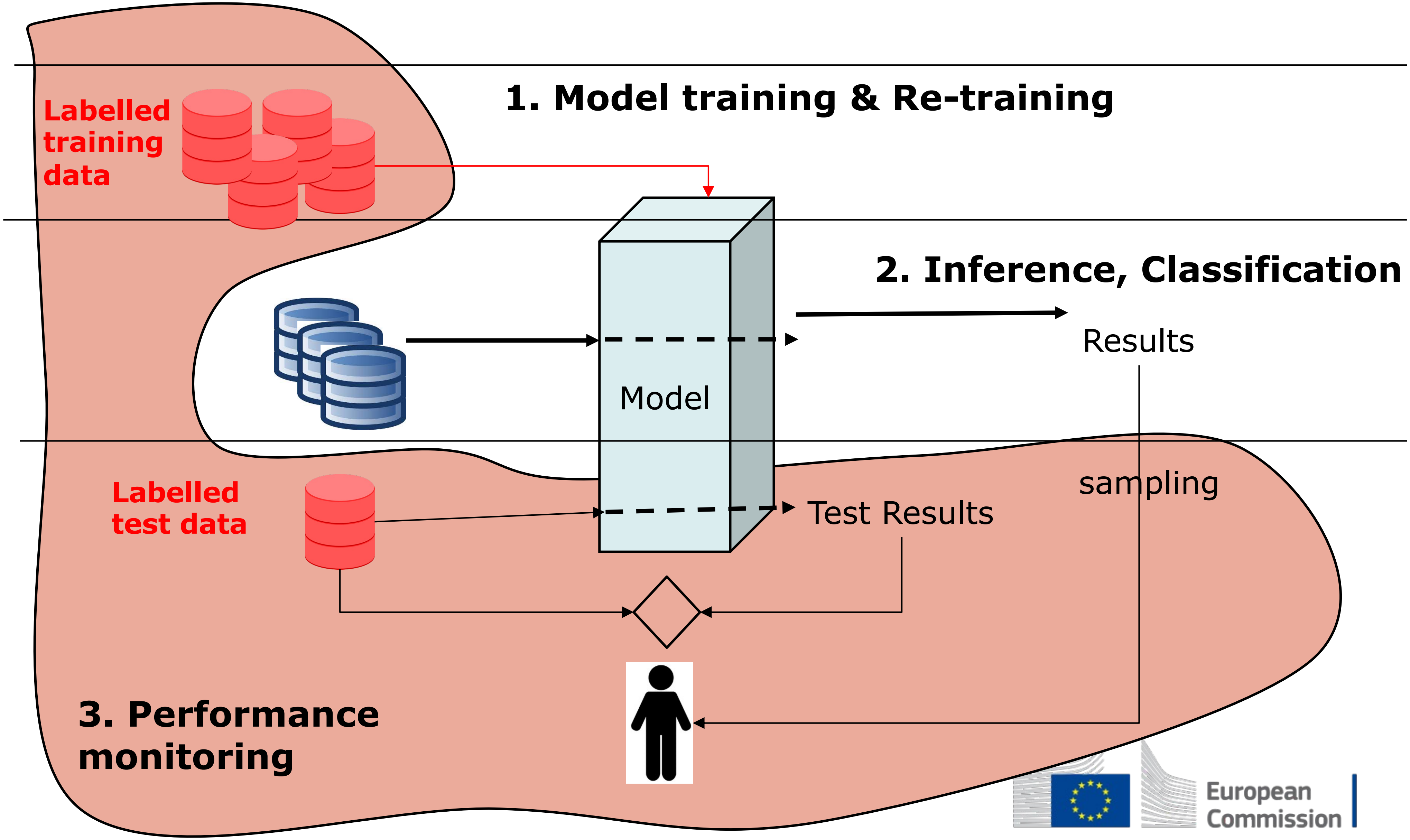
eurostat

EG B2G4S

A few words on Quality aspects of supervised ML in OS


Labelled training data
1. Model training & Re-training
2. Inference, Classification
Results
Model
sampling
Labelled test data
Test Results
3. Performance monitoring

A few words on Quality aspects of supervised ML in OS



1. Model training & Re-training

2. Inference, Classification

3. Performance monitoring

A few words on Quality aspects of supervised ML in OS

- ***Quality framework must cover processes and data related to (re)training and monitoring***
  - Quality of traning data → quality of ML model → quality of ML results

- *Examples of issues to be addressed (guidelines, criteria)*
  - How do you assess/monitor/audit the performances of the ML? Does that involve human inspection?
  - What conditions/criteria will indicate that a fresh re-training is needed? E.g., drift of performance metrics?
  - Provenance of training data: How do you produce training data? Who labels them? How many people and with what skills will be in charge of labelling? (NB: there are cost-accuracy tradeoffs)
  - Can (or should) you publish the training data for the sake of methodological transparency?
  - Versioning ML models; versioninsg of training data
  - Is the energy consumption significant at any stage?
  - …

Conclusions

- *Transition from* ***experimental statistics*** *to* ***official statistics*** *requires walking the whole quality path*

- *Developments in* ***methodologies*** *and* ***quality aspects*** *must go hand-in-hand*

- *Mind the strategic implications of methodology/quality choices, e.g. NSI vs data holder(s)*

- *Concrete guidelines and operational criteria must be specific to data classes, but advancements on one data class (e.g., MNO data) will be useful (or at least inspirational) for other classes*

Thanks for your attention

Quality Assurance Framework of the European Statistical System

EUROPEAN STATISTICAL SYSTEM

Version 2.0