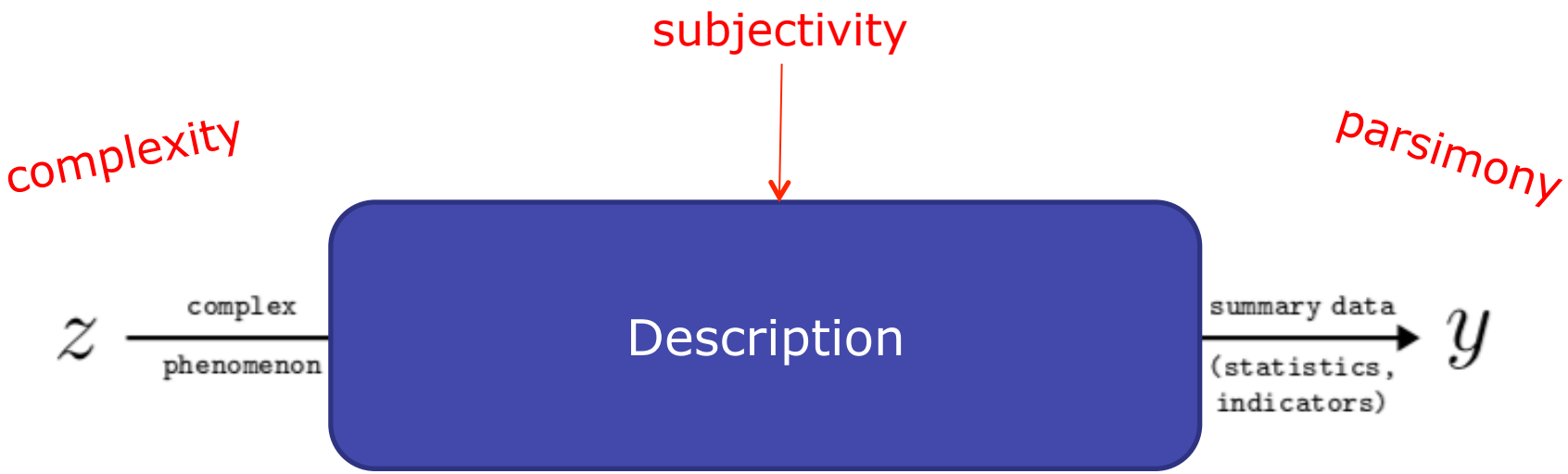
Public manuals and open-source code: rethinking rethodological documentation for new data sources

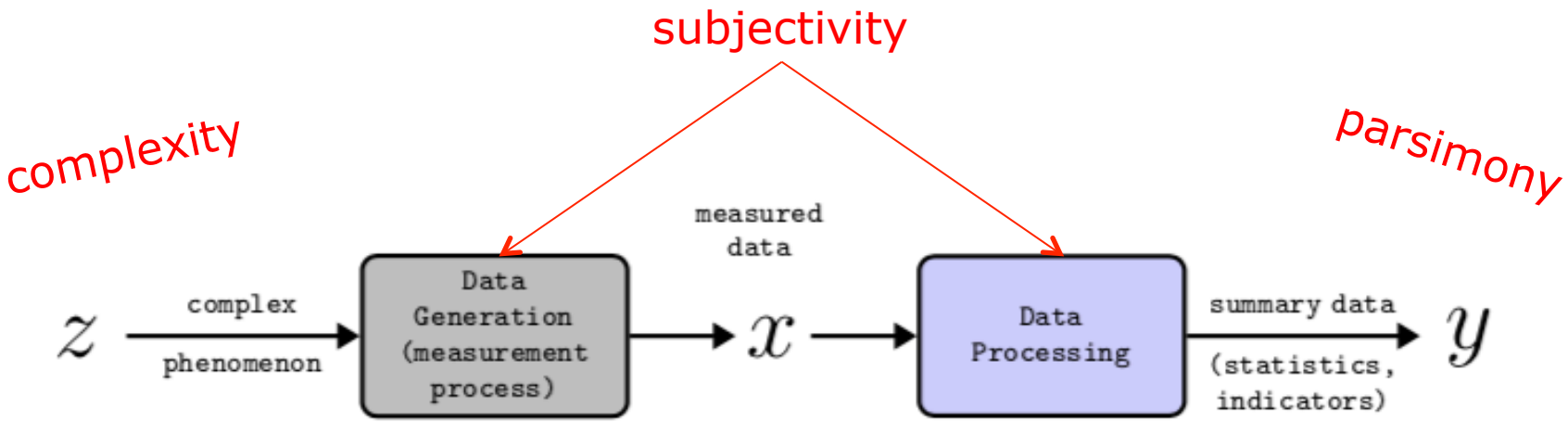
Fabio Ricciato
Jacopo Grazzini
Jean-Marc Museux

European Commission, Eurostat
Unit B1 – Methodology; Innovation in Official Statistics

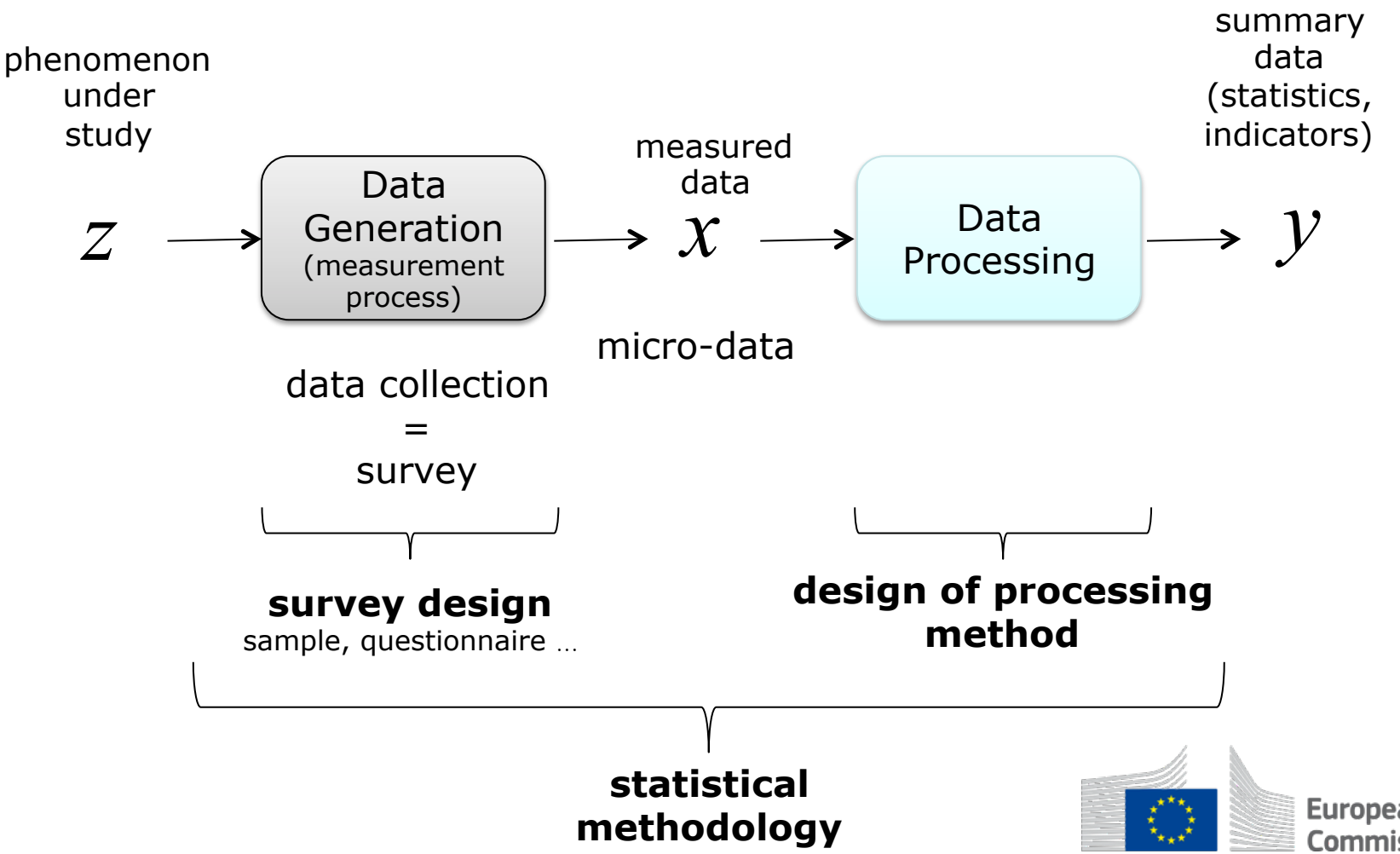
Statistics = description through a measurement process



Statistics = description through a measurement process

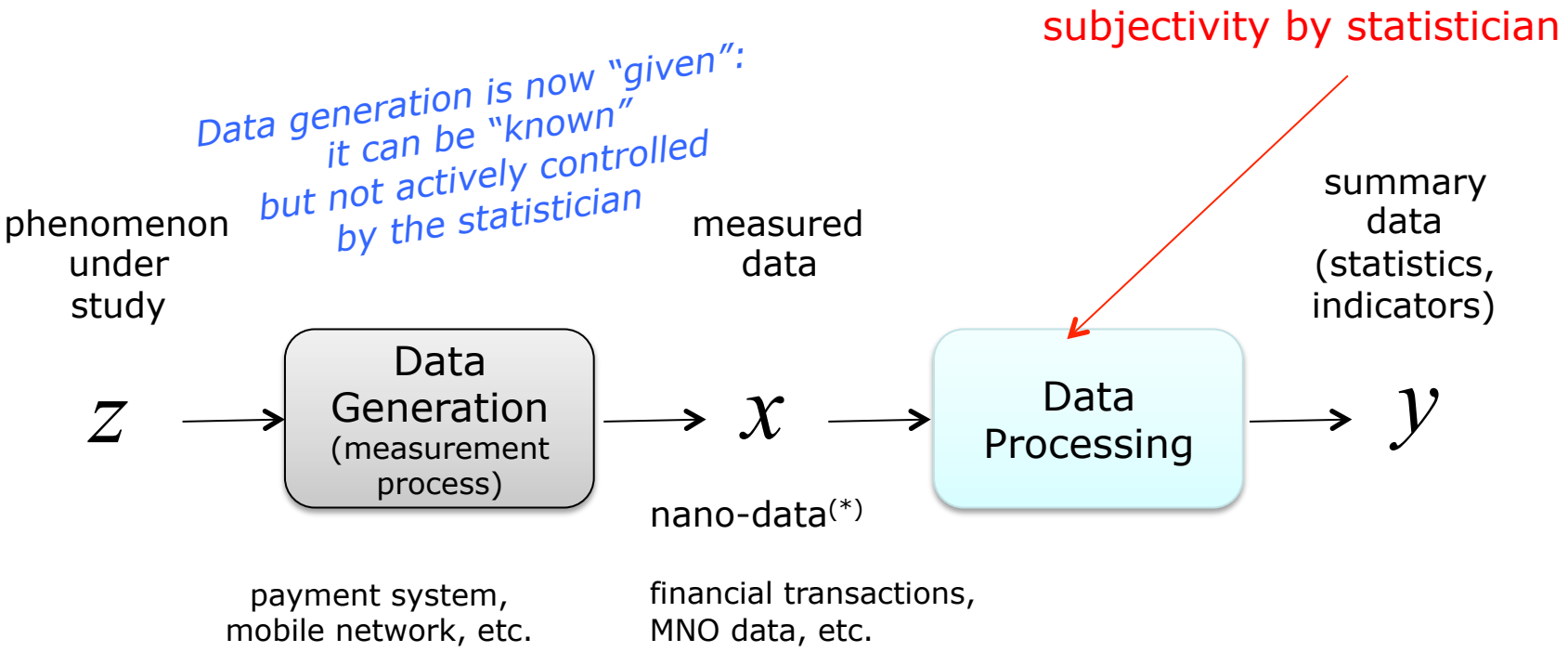


Statistical methodology for survey data



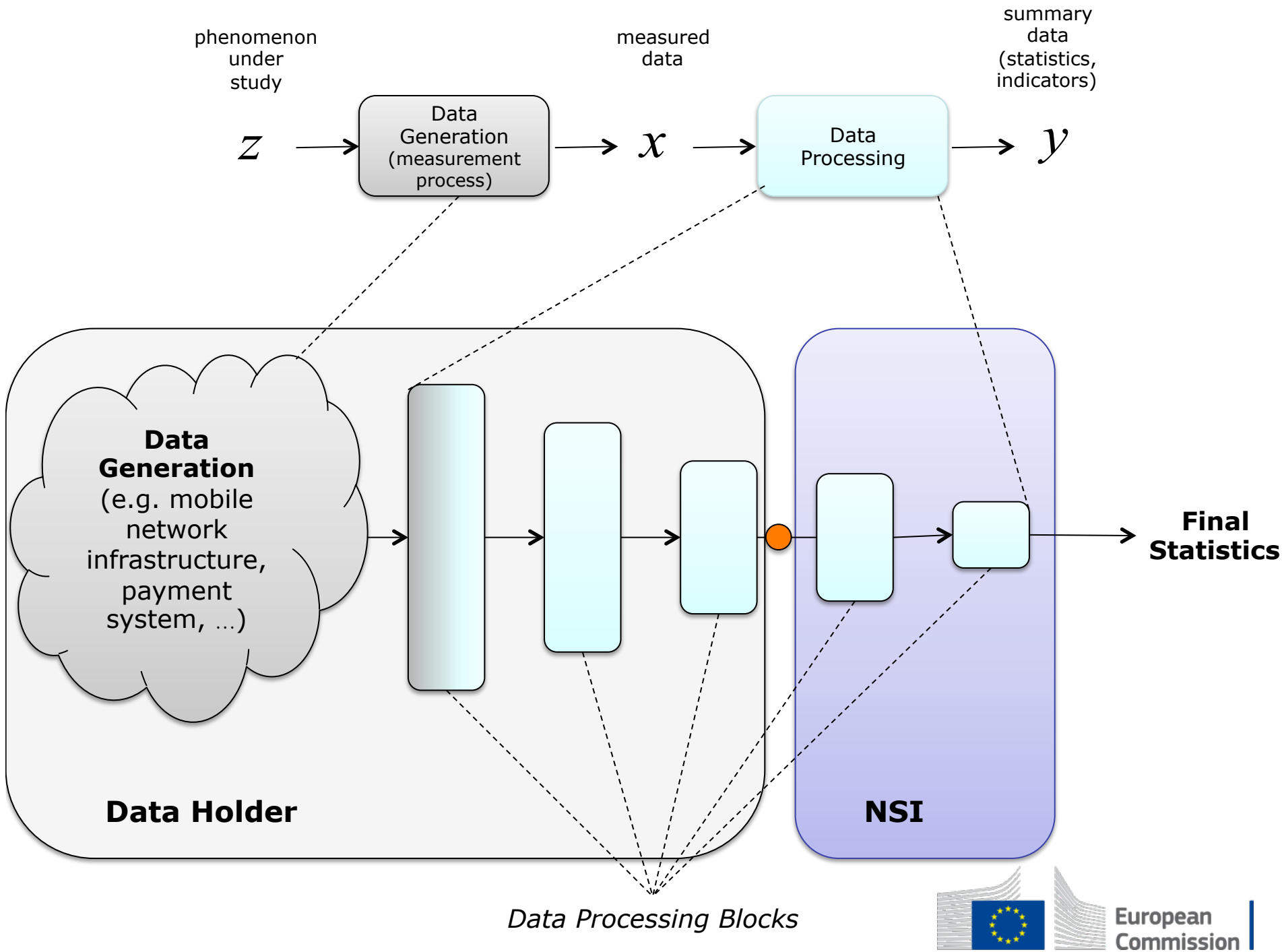
European Commission

Statistical methodology for "given data"



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

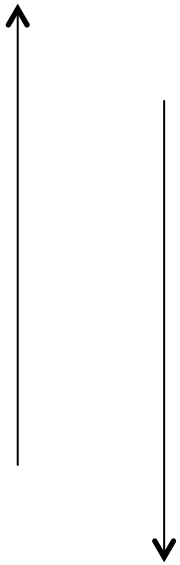
European Commission



European Commission

Levels of description

Understandable,
Explainable



Reproducible
Modifiable

Human-readable manuals

Pseudo-code

(open) source code
+ code documentation
+ benchmark data

Tue, 9 Mar | Wed, 10 Mar | Thu, 11 Mar

11:00 AM - 12:00 PM [Europe/Brussels] We-STS03
User demand during uncertain times

Back

Statistics Coded – Storytelling through literate programming and runnable computing

Authors et al., **(Email)** [2], Authors et al., **(Email)** [3], J. Davis, **(Email)** [1], H. Lehtimaki, **(Email)** [1], M. Meszaros, **(Email)** [1], **J. Grazzini, (Email)** [1], et al.

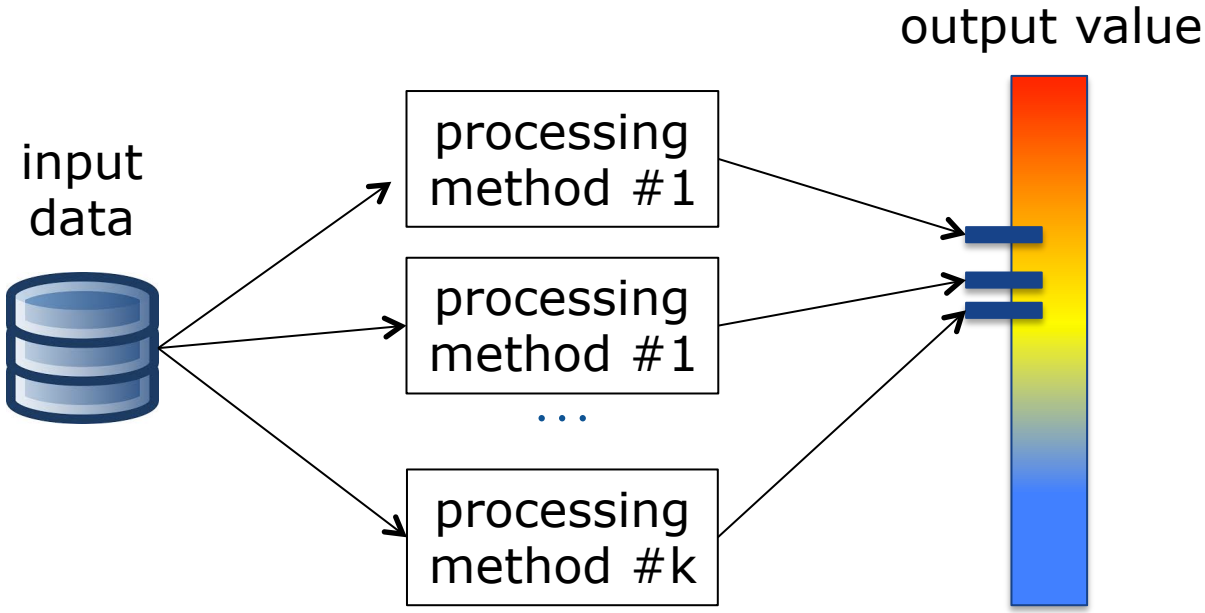
[1] *Eurostat - European Commission*
[2] *European Master in Official Statistics*
[3] *Eurostat Blue Book internship program*

This paper introduces the so-called *Statistics Coded* as a way to disseminate outputs that document and share statistical processes beyond just data, but also including code, algorithms, protocols and workflows. They enable users interactively dialoguing with data, explore, reuse and contribute to statistical products. They help them as well sharing best practices, learning from each other, and adopt common methodologies, fostering *Statistical Literacy* in the public. Inspired by the *Open Source Software*, *Citizen Science* and *Open Science* communities, *Statistics Coded* present substantial promises for *Citizen Statistics* and the interaction between *National Statistical Institutes*, data users and data producers.

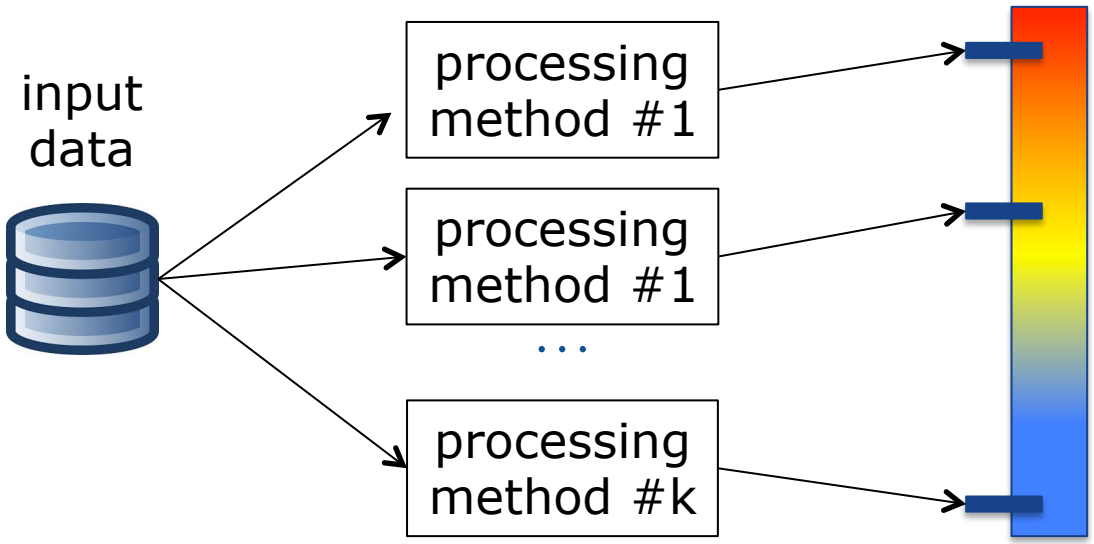
Statistics Coded – Storytelling through literate programming and runnable computing
https://coms.events/NTTS2021/data/abstracts/en/abstract_0041.html

European Commission

Methodological sensitivity



Low Methodological Sensitivity

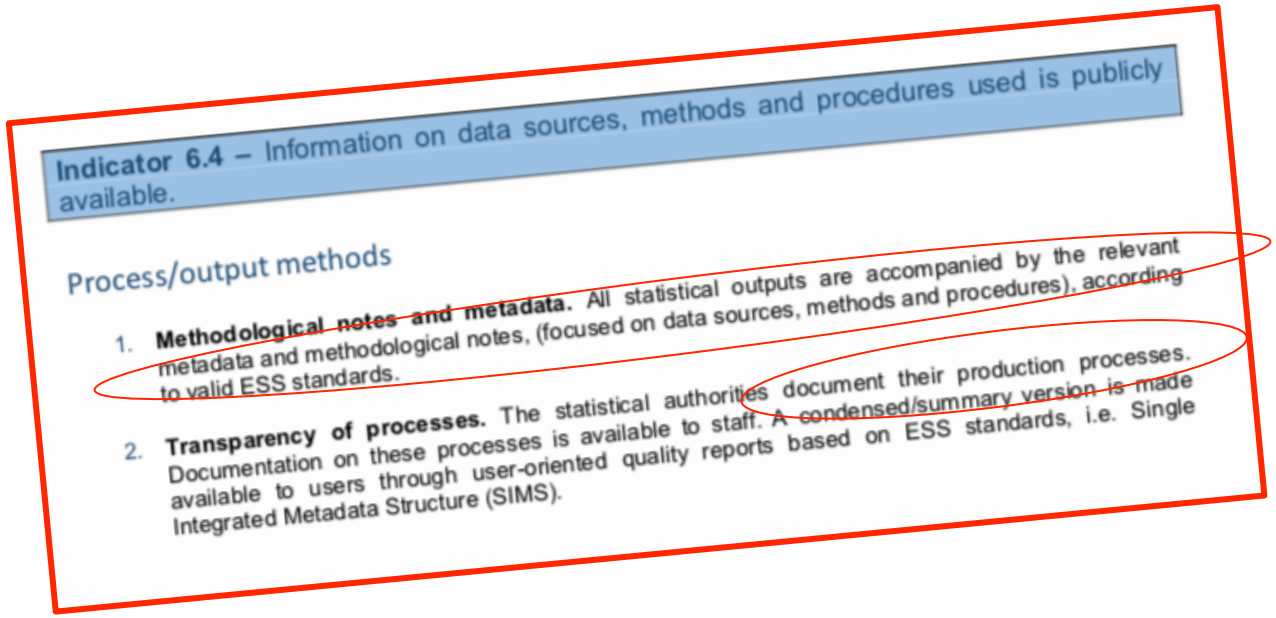


High Methodological Sensitivity

European Commission

About methodological development

- *Goals of methodological development*
  - **Produce** high-quality methods and evolve them as needed (→ continuous process, not one-off task)
  - **Build trust** in the methods: how the methods and the methodological development process itself are documented, communicated, audited, etc. matters!
  - **HOW** methods/results are produced matters no less than **WHAT** methods/results are produced



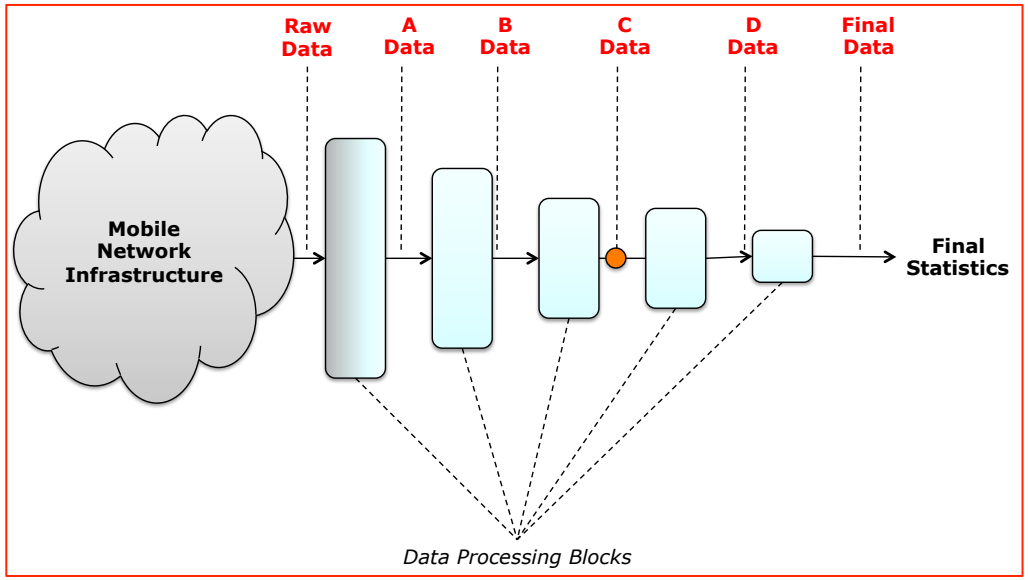
Indicator 6.4 – Information on data sources, methods and procedures used is publicly available.

Process/output methods

1. **Methodological notes and metadata.** All statistical outputs are accompanied by the relevant metadata and methodological notes, (focused on data sources, methods and procedures), according to valid ESS standards.

2. **Transparency of processes.** The statistical authorities document their production processes. Documentation on these processes is available to staff. A condensed/summary version is made available to users through user-oriented quality reports based on ESS standards, i.e. Single Integrated Metadata Structure (SIMS).

European Commission

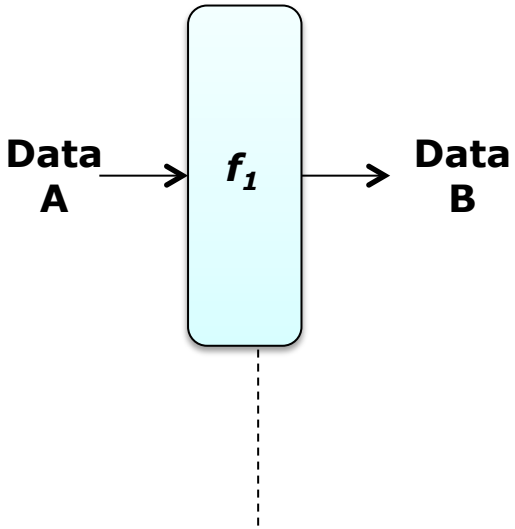


The description of method f_1 is "meta-data" for data B

meta-data for A
+ method description f1
= meta-data for B

$$B = f_1(A)$$

$$\langle A, f_1 \rangle \rightarrow B$$



Data Processing Block implementing method $f_1()$

European Commission

Summary

- *New data are big and complex to interpret*
  - Data processing executed by machines, methodology becomes "code"
- *Processing methodologies tend to be more complex, with fatter pipelines richer of design choices, parameters, …*
  - A modular design is key! As in complex software projects
- *Methodological sensitivity increases*
- *Methodological documentation limited to human-readable manuals & guidelines is not sufficient*
- *Regular methodological documentation must be augmented to include open-source code*
  - + code documentation, + benchmark data
- *Modular structure + complete documentation …*
  - Enables reproducibility, auditability –> brings credibility, trust!
  - Enables collaborative development → lower costs
  - Facilitates adaptation and evolution → better quality

Thanks for your attention

For follow-up contact:

fabio.ricciato@ec.europa.eu