



Trusted Smart Statistics: motivations and principles

Fabio Ricciato
European Commission
Eurostat

*ECA conference on Big and Open Data
for European Union supreme audit institutions
Luxembourg 28.11.2019*

What “Big Data” means in Official Statistics?

- Surveys and census – asking the people
- Administrative records – what people declared to public administrations

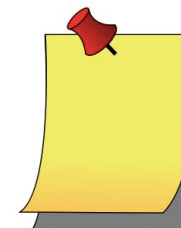
traditional
data

- “Big Data” - everything else

- Satellite images
- Credit Card transactions
- Mobile Network Operator data
- Web, Online Social networks (Internet-as-a-data source)
- Traffic sensors, street cameras
- Air Traffic Control Data, vessel sheep
- Smart meters
- ...

non-traditional data
aka
new digital data

Key point #1

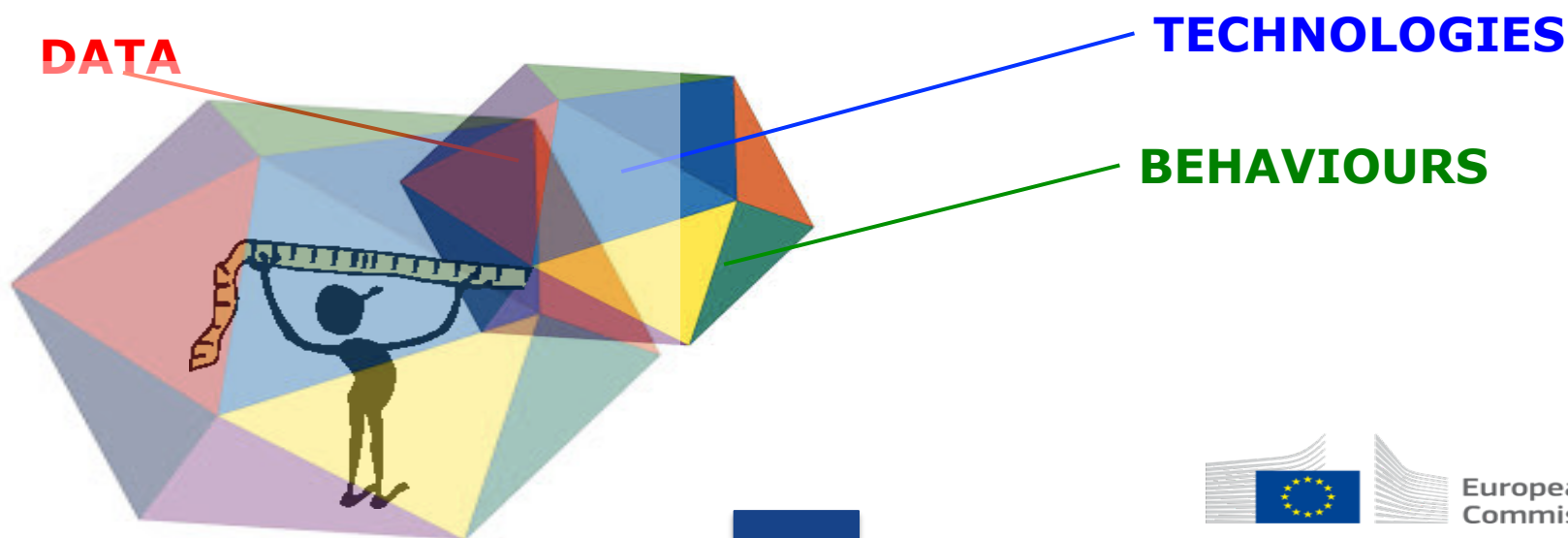


- What matters most is *not the size* (quantitative) but *THE CHARACTERISTICS* (qualitative) of new data
- What matters most is not that they are more/bigger but that they are *different* (from traditional data sources)

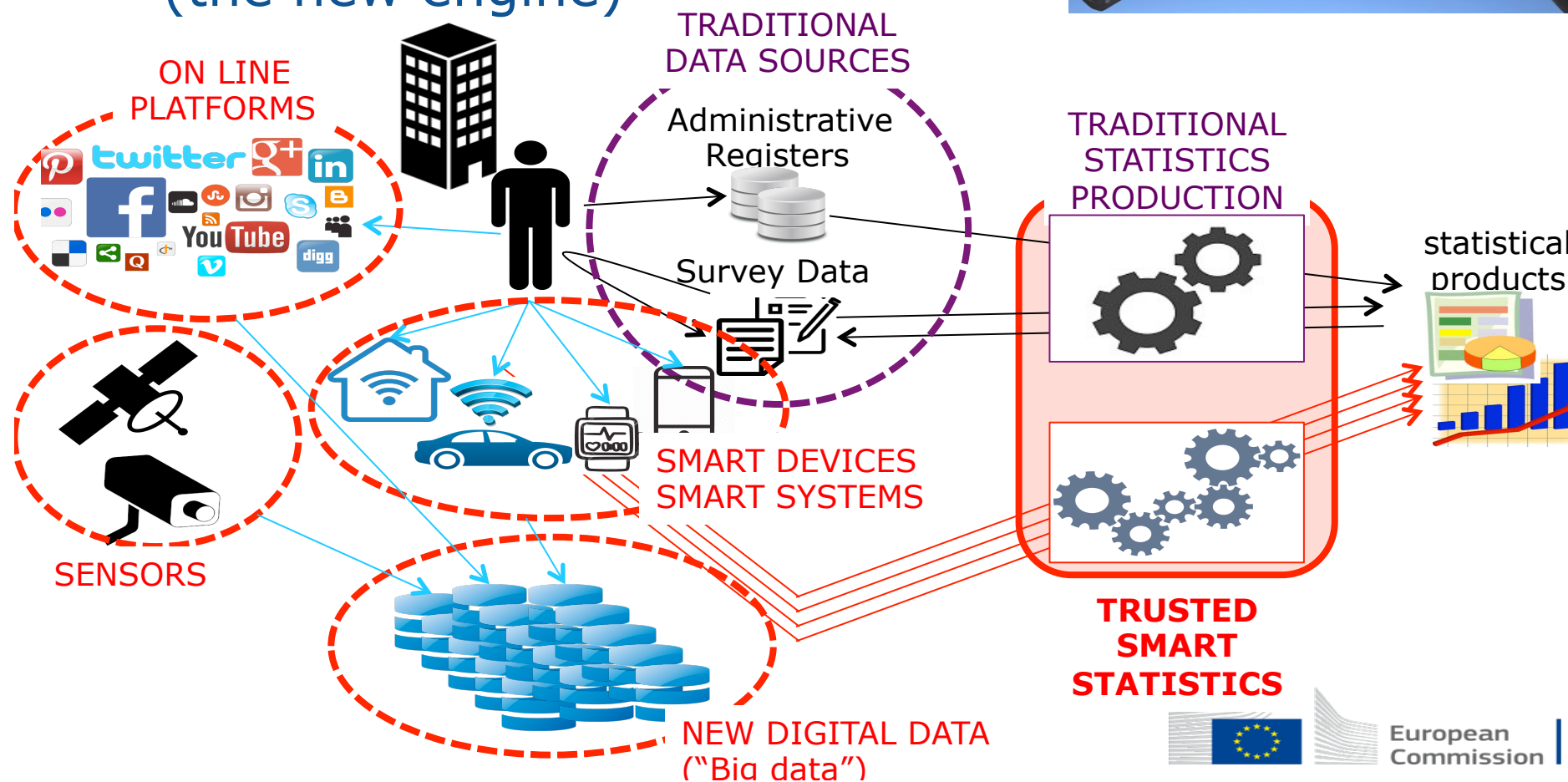


Key point #2

- New digital **data** come with new digital **technologies** and new digital **behaviours** and perceptions, attitudes, expectations ...
- It's a new digital world (new data one of its facets)



Trusted Smart Statistics (the new engine)



Designing the new engine

Trusted Smart Statistics (TSS) → systemic augmentation of official statistics

take a system-level view
define a clear “grand picture” first, then
develop components based on that...

the new processes
add to / integrate with
legacy ones.

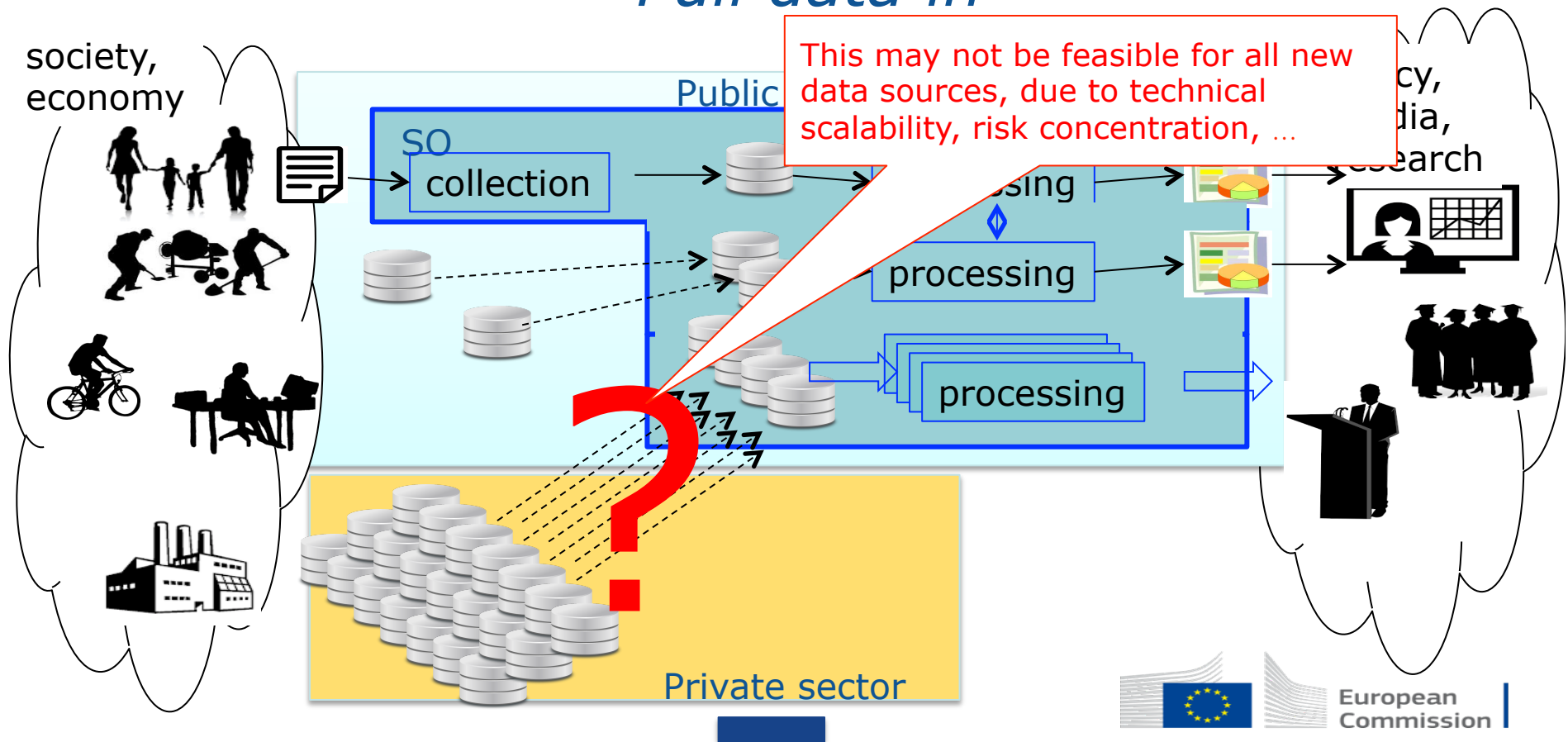
A solid development starts from a solid design.

A solid design starts from clear **design principles**



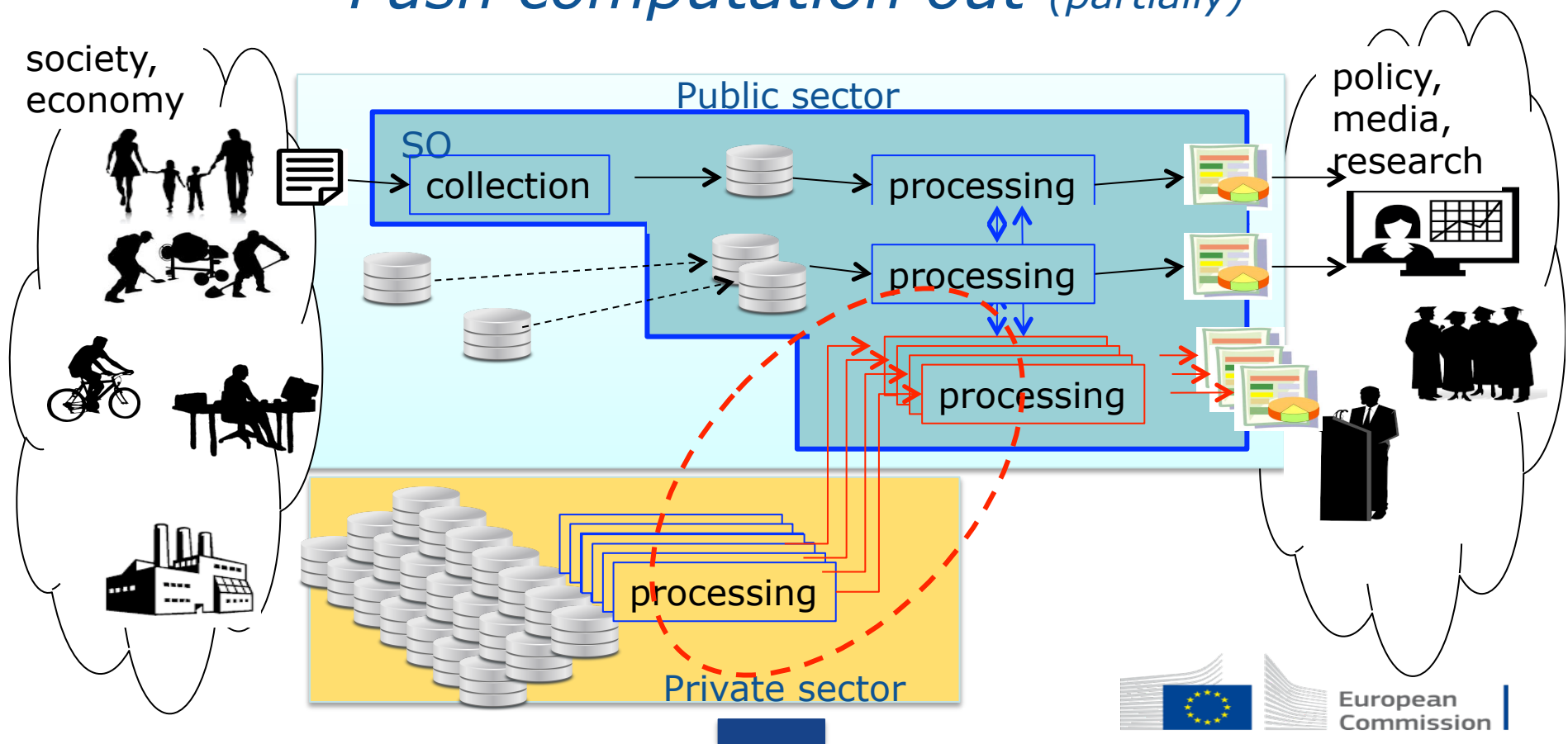
Handling the new in the old way

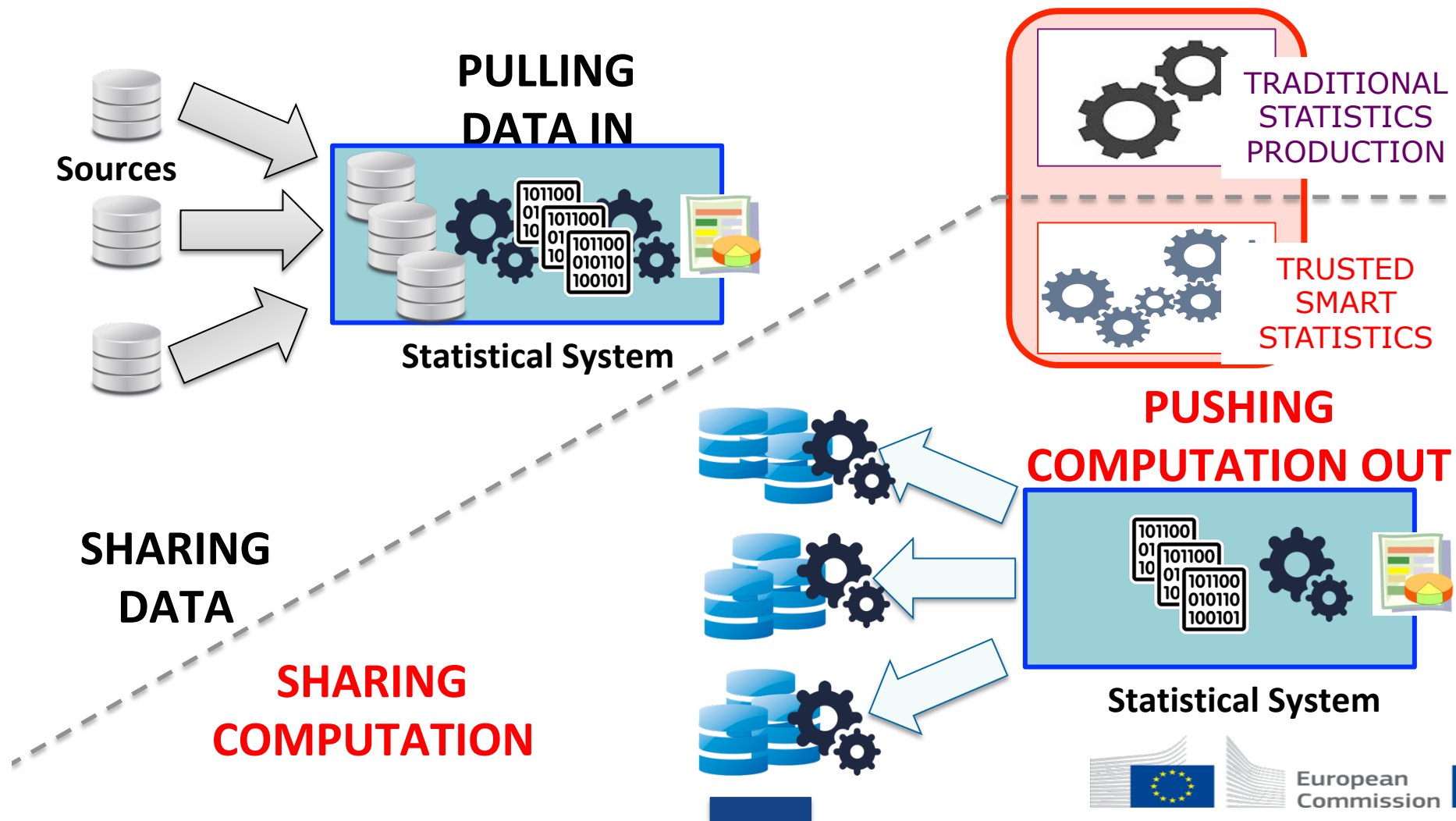
Pull data in



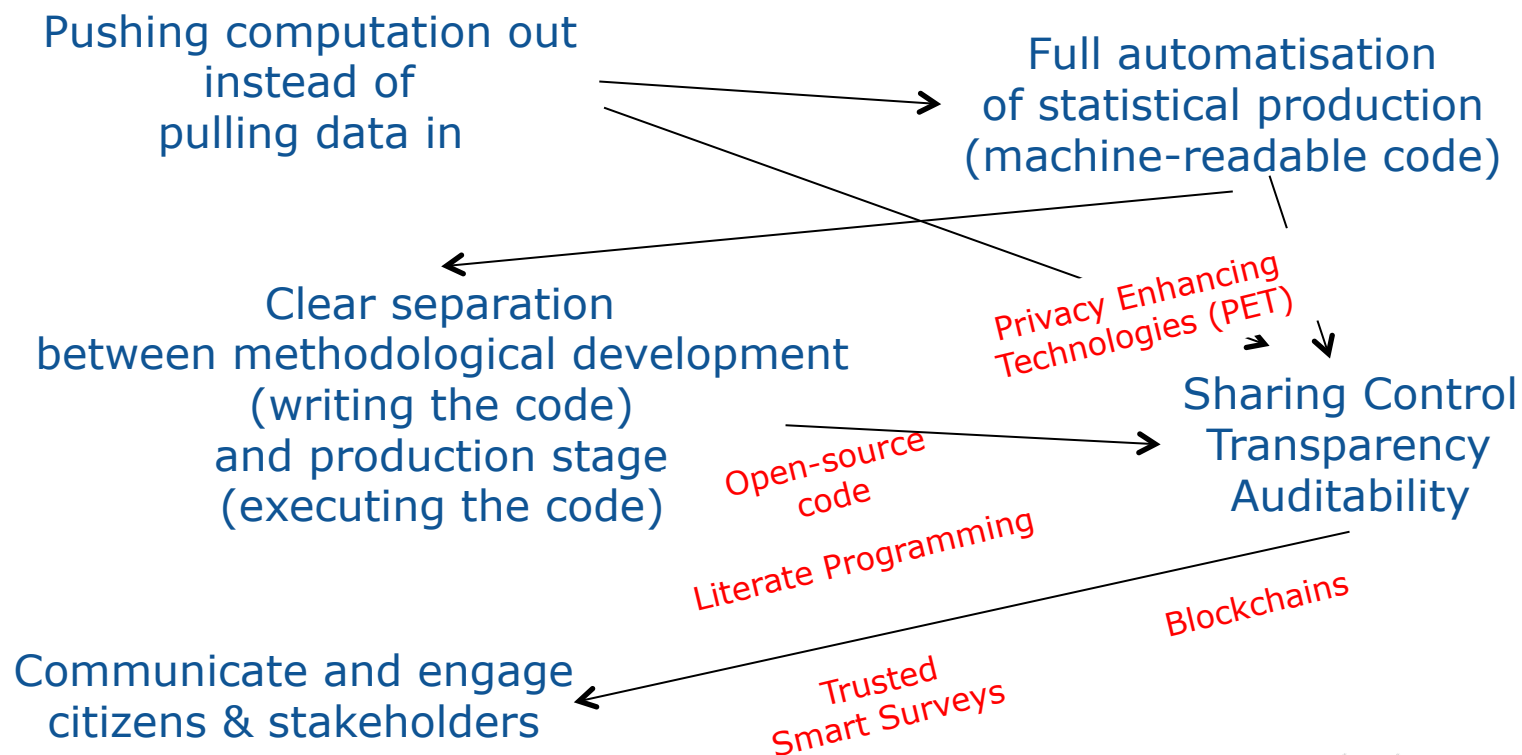
Handle the new in new ways

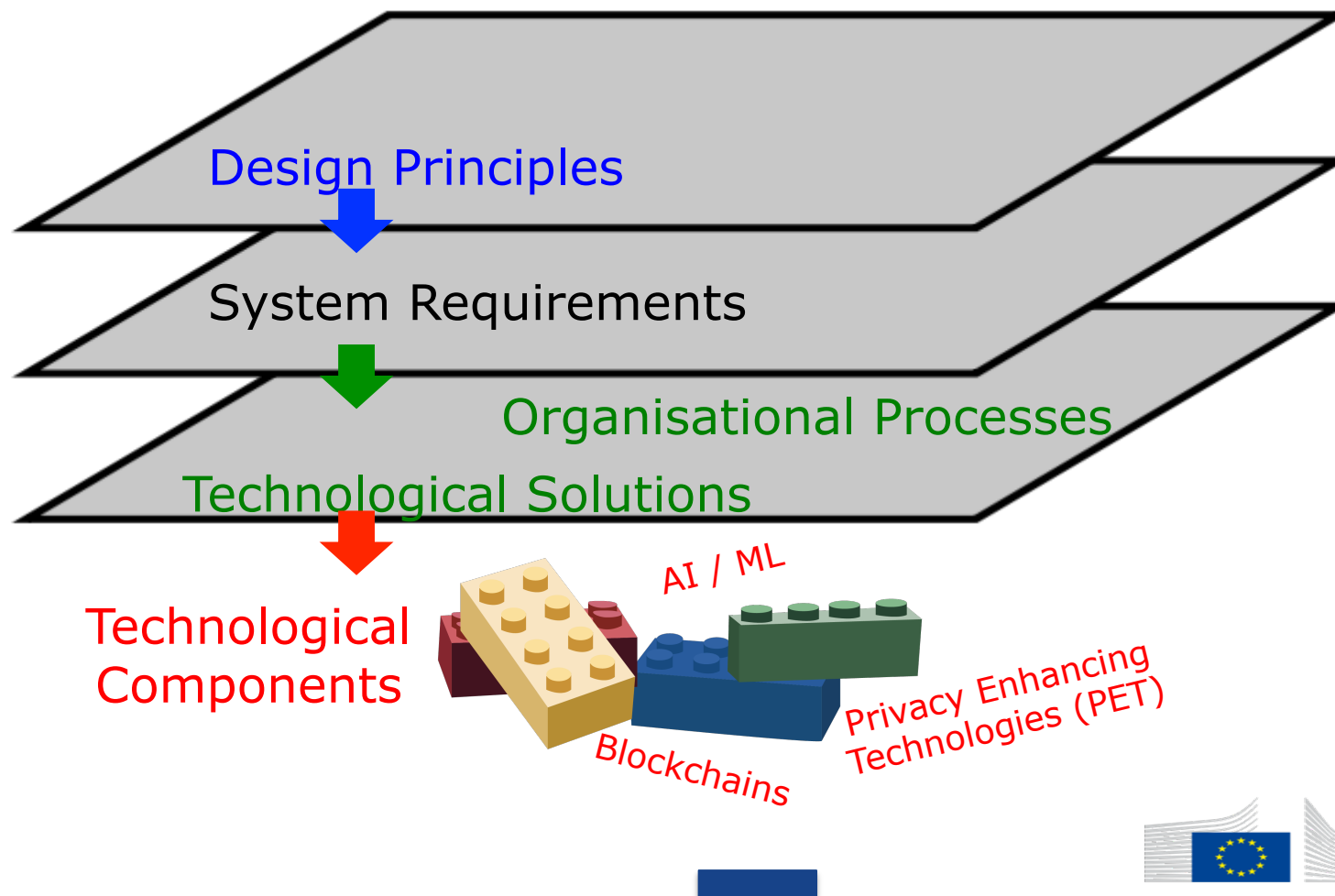
Push computation out (partially)





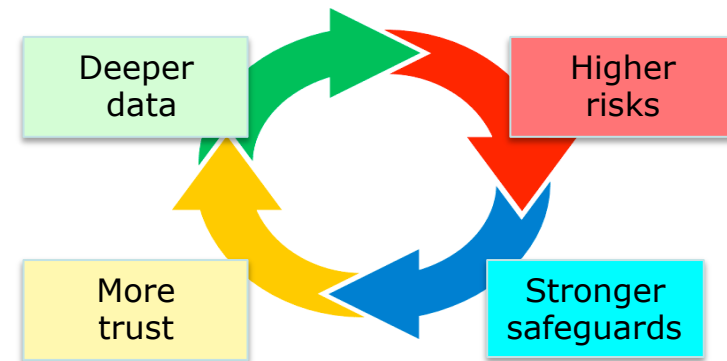
Design principles for TSS





Transparency and public trust

- Close the (confidential) data, open the algorithms.
- Share control, share logs, share algorithms.
Don't share confidential data.
- Goal: ensure shared, **public control** on ***how the data are used.***
 - Which queries are run on the data? By whom? For what purpose?
By which analytic methods?
- Transparency and auditability
 - Always publish the source code.
Immutable logging of queries
 - Privacy Enhancing Technologies



Focus on methodological development



Analog World

- Scarce data: costly to collect
- Designed data: easy to interpret
- Slow change: time to consolidate methodologies



Digital & Datafied World

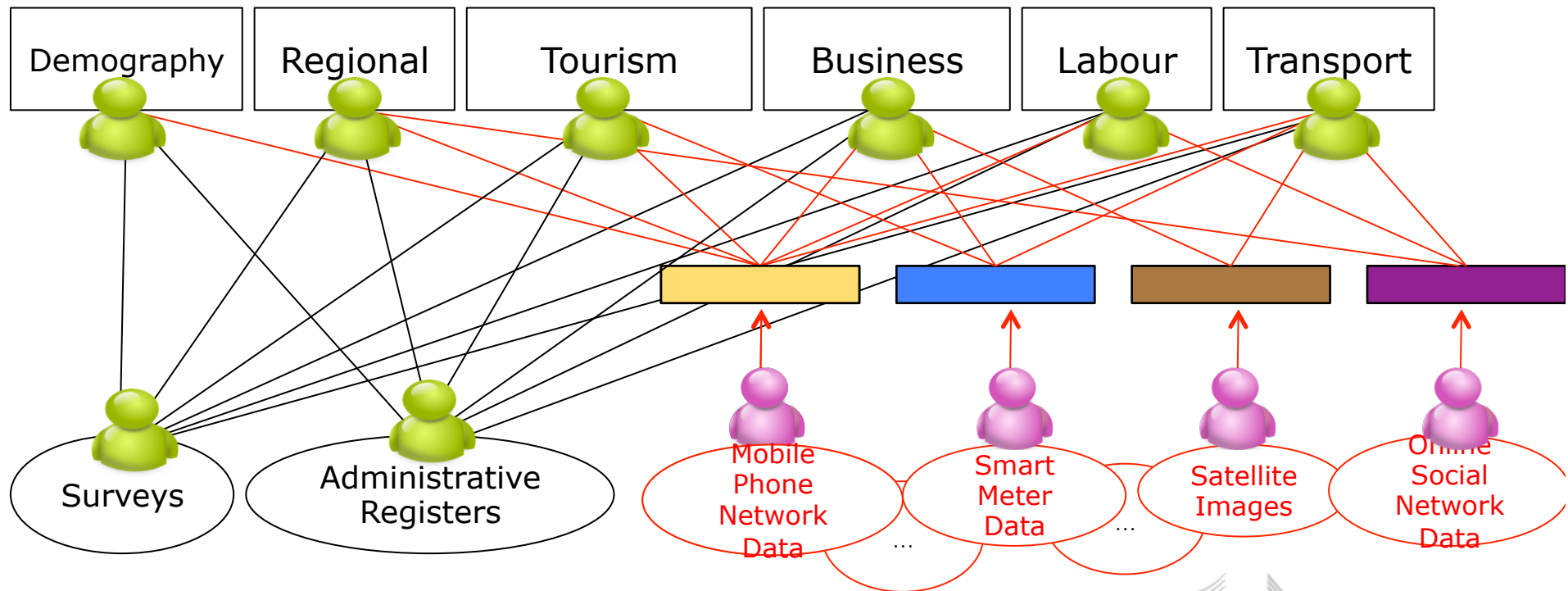
- Abundant data: already there (collected by somebody else)
- Found data: difficult to interpret
- Fast change: pressure on methodological developments



European
Commission

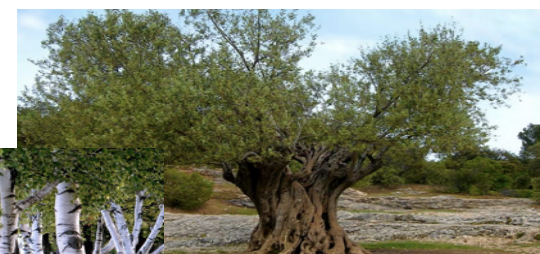
Multipurpose Data & Multisource Statistics

Statistical domains

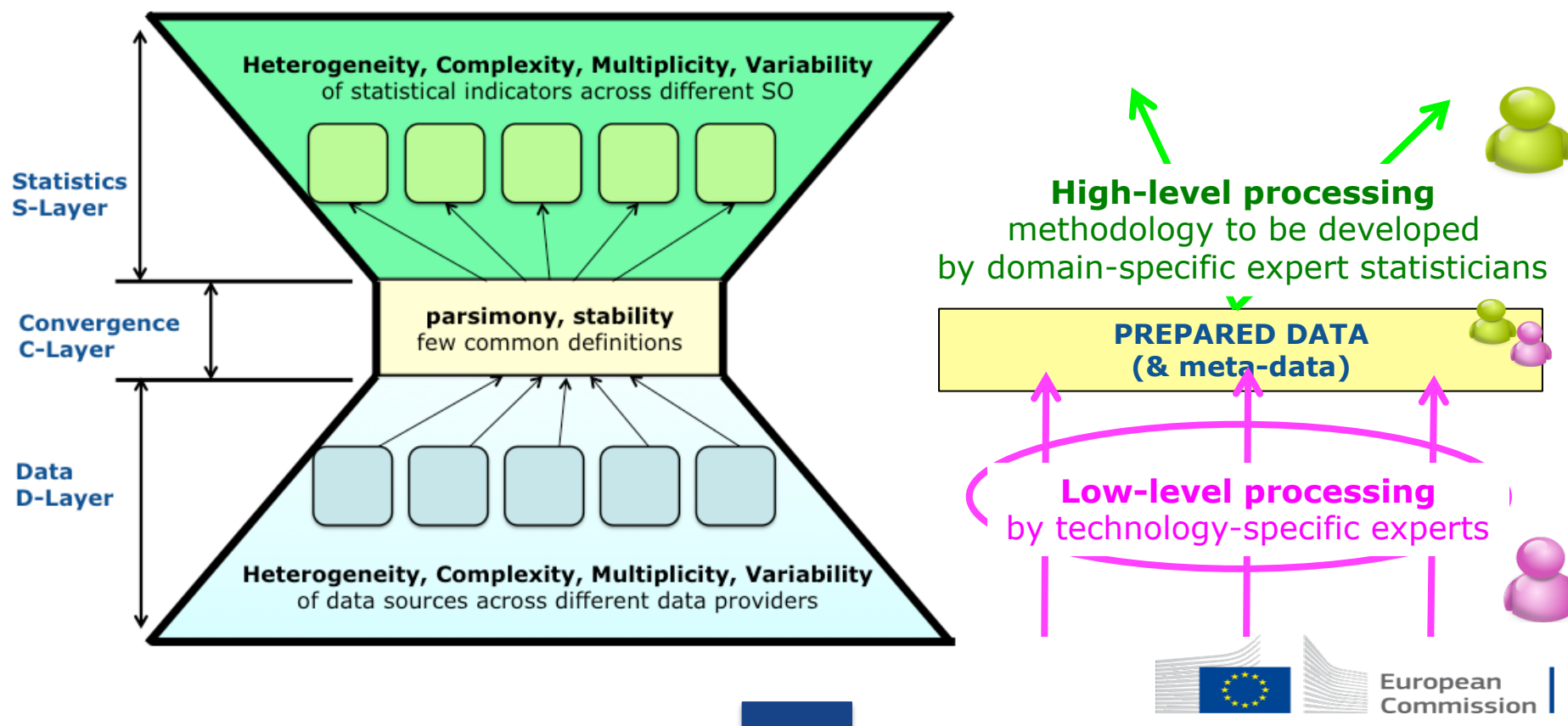


Classes of Sources

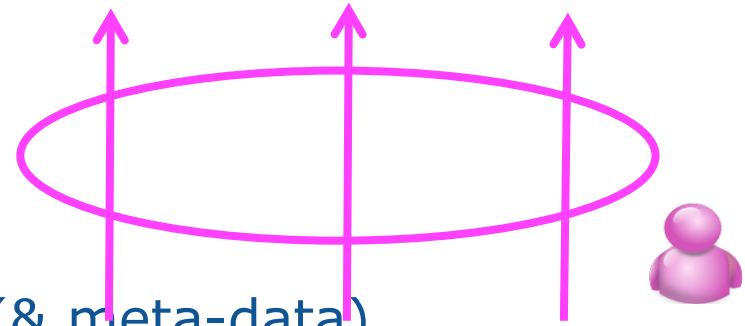
Multipurpose Data & Multisource Statistics



Hourglass model

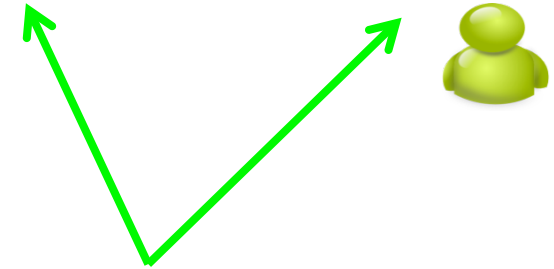


Challenges at the bottom



- preparation of intermediate data (& meta-data)
 - the **lower** part of the methodology workflow
 - transforms complex technology-heavy raw data into “statisticians friendly” intermediate data
 - requires source-specific skills, not domain-specific
- data access
 - voluntary agreements with data holders is sufficient for experimental statistics and pilots
 - guaranteeing sustainable access (as needed for regular production) might require new legislation

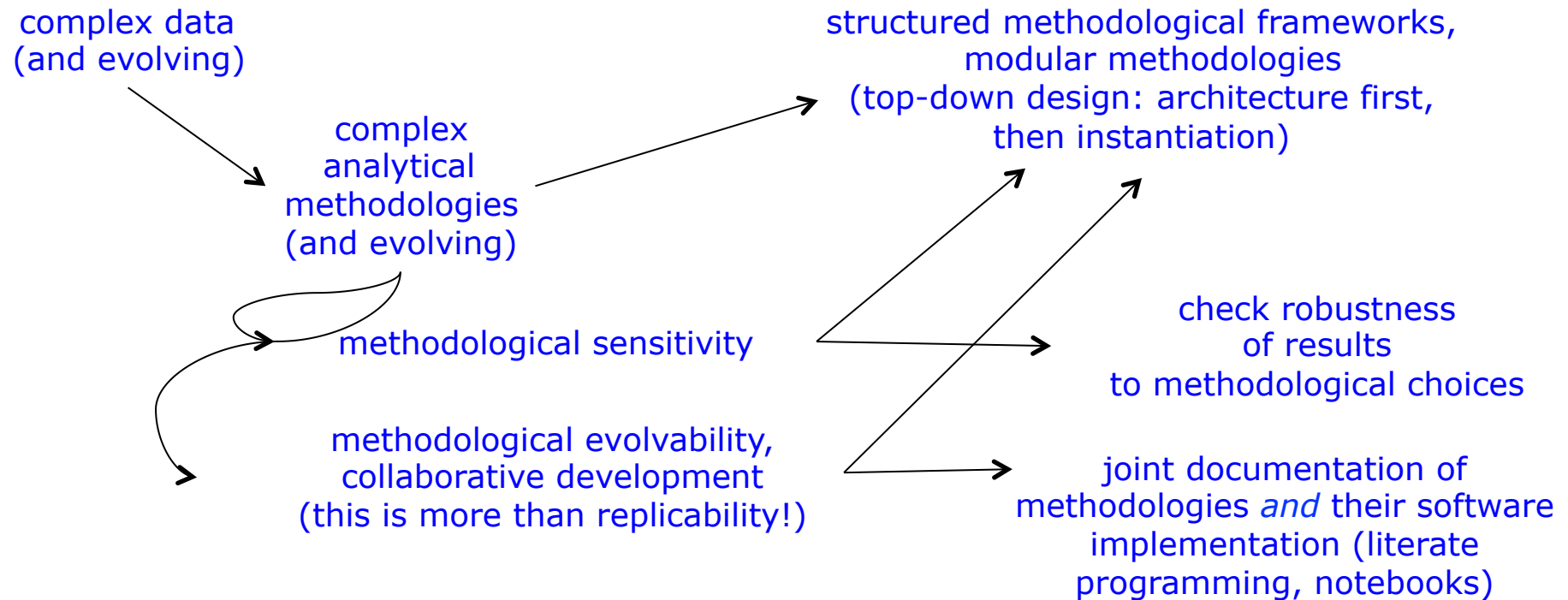
Challenges at the top



- statistical processing of prepared data (& meta-data)
 - the **upper** part of the methodology workflow
 - transforms “statisticians friendly” intermediate data into final indicators, or (experimental) statistics
 - requires domain-specific skills → production units
- integration with legacy data / into legacy statistics
 - fusing new data sources and survey/admin data
 - incorporating (information from) new data sources into legacy statistics

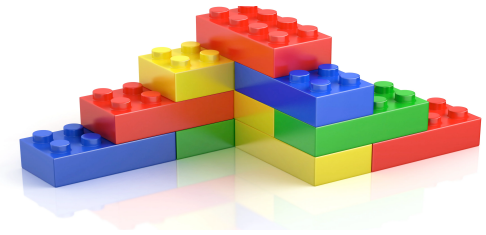


Complex data → Complex analytics



... like with complex software
development ...

Take home messages



- Using new data sources for official statistics requires a fresh system-level view and new approaches about ...
 - how data are accessed (trust engineering)
 - how analytical methodologies are developed and evolved (collaboratively)
- As analytical methodologies gets represented as complex software, we should learn tools and concepts from software architects (modularity, modularity, modularity...)
- New technologies (PET, blockchains, ...) may play on our side, but we must understand them and learn how to combine them (they are tools, not goals)
- “Big data” is more about people and processes than about data





Thanks for your attention

for follow-up: *fabio.ricciato@ec.europa.eu*