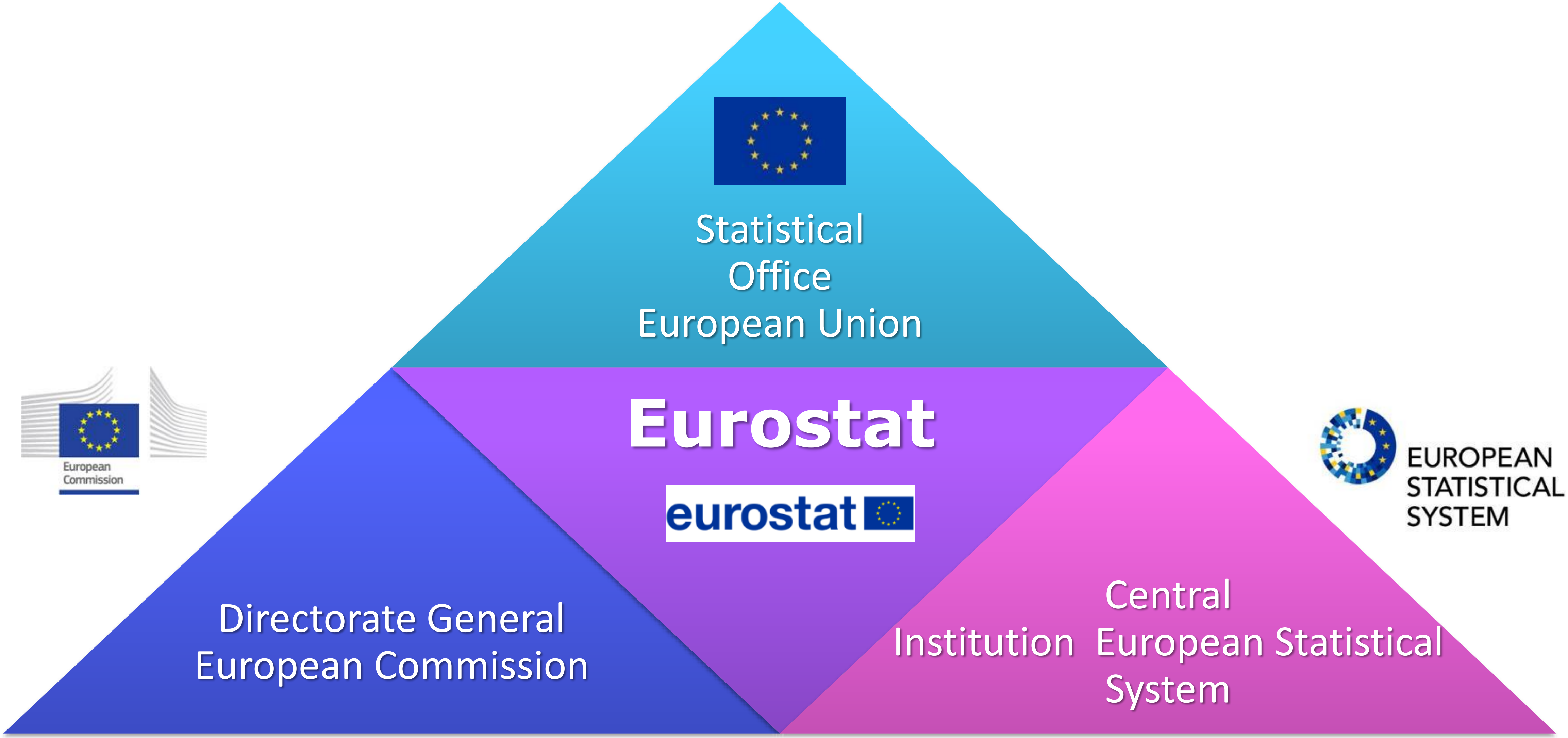
Trusted Smart Statistics

EMOS webinar

Albrecht.Wirthmann@ec.europa.eu

Luxembourg, 25 Feb 2020



Statistical Office European Union

Eurostat

eurostat

Directorate General European Commission

Central Institution European Statistical System

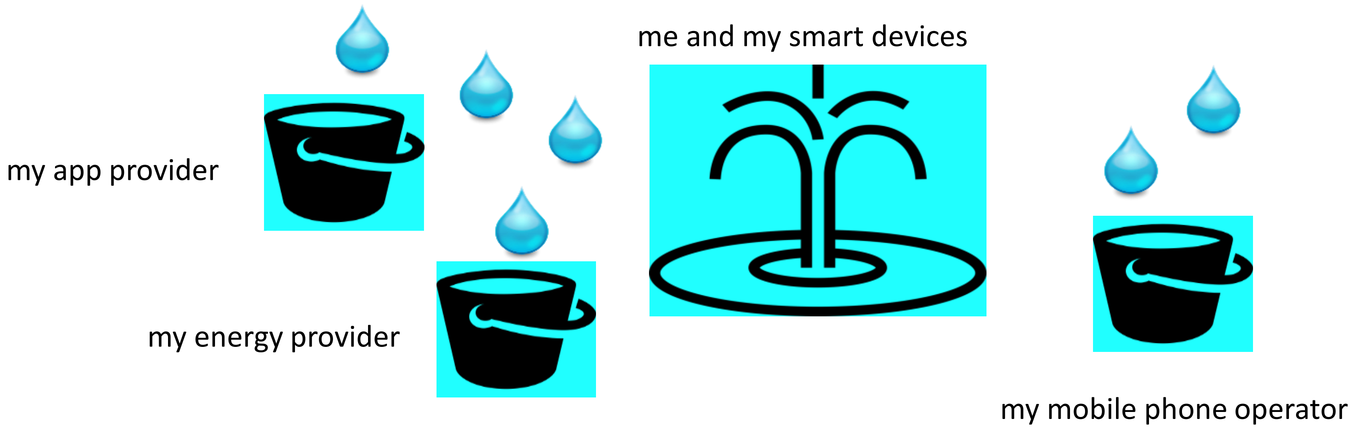
EUROPEAN STATISTICAL SYSTEM

The new datafied world

- “**Anything that goes digital, gets logged**”
  (somewhere, by somebody) 1° fundamental law of datafication

digital transformation → datafication

- Individuals, organizations, places … become “data **fountains**”
- More and more business companies become “data **buckets**”



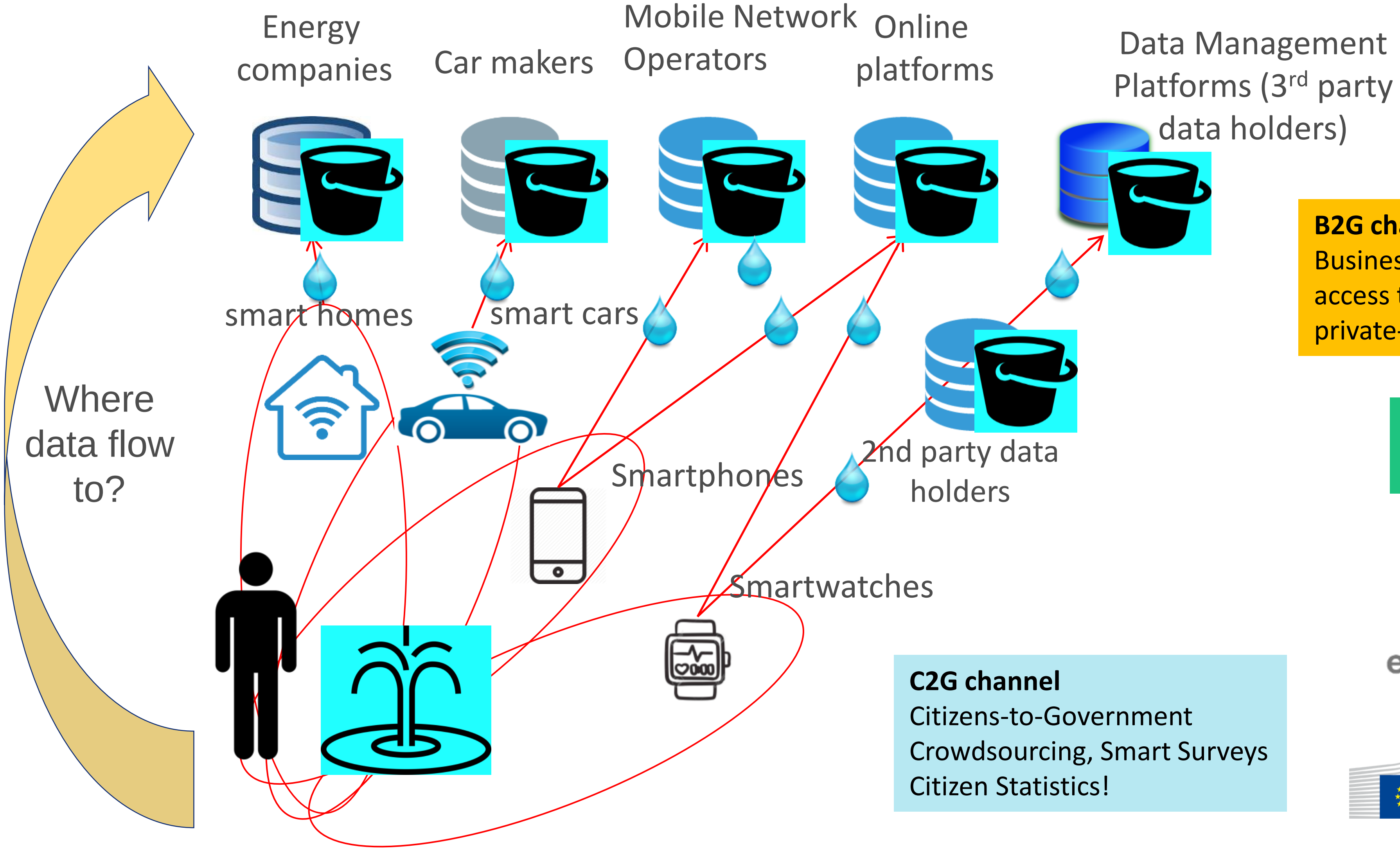
me and my smart devices

my app provider

my energy provider

my mobile phone operator

Fountains or from buckets?



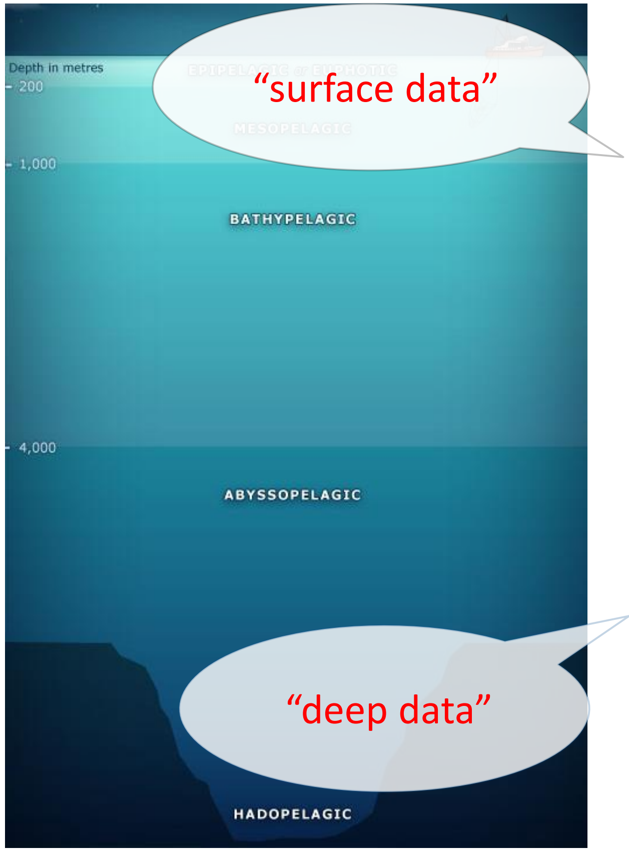
B2G channel
Business-to-Government
access to privately-held data
private-public partnerships

Statistical Office

eurostat

C2G channel
Citizens-to-Government
Crowdsourcing, Smart Surveys
Citizen Statistics!

Surface data and deep data



“surface data”

“deep data”

“micro-data”

- Name, gender, date of birth
- 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 heartbeat, blood pressure…
- Every single transaction, events involving you …

…

Highly pervasive data on features changing constantly and recorded at fine timescale

Implications for data access,
data /process governance, privacy and
confidentiality

Traditional Sources

(survey/census, admin records)

- Micro-data
  individual level
- Designed data,
  purposed for OS
  or for admin. process
- Structured
- Always collected within **Public Sector** institutions
- …

New Data Sources

(all others)

- Nano-data
  sub-individual level
- Organic data,
  re-purposed
  for Official Statistics
- Structured, semi-structured, unstructured
- Often held by
  Private Sector companies
- …

Key point #1

- What matters most is *not the size* (quantitative) but *their characteristics* (qualitative) of new data

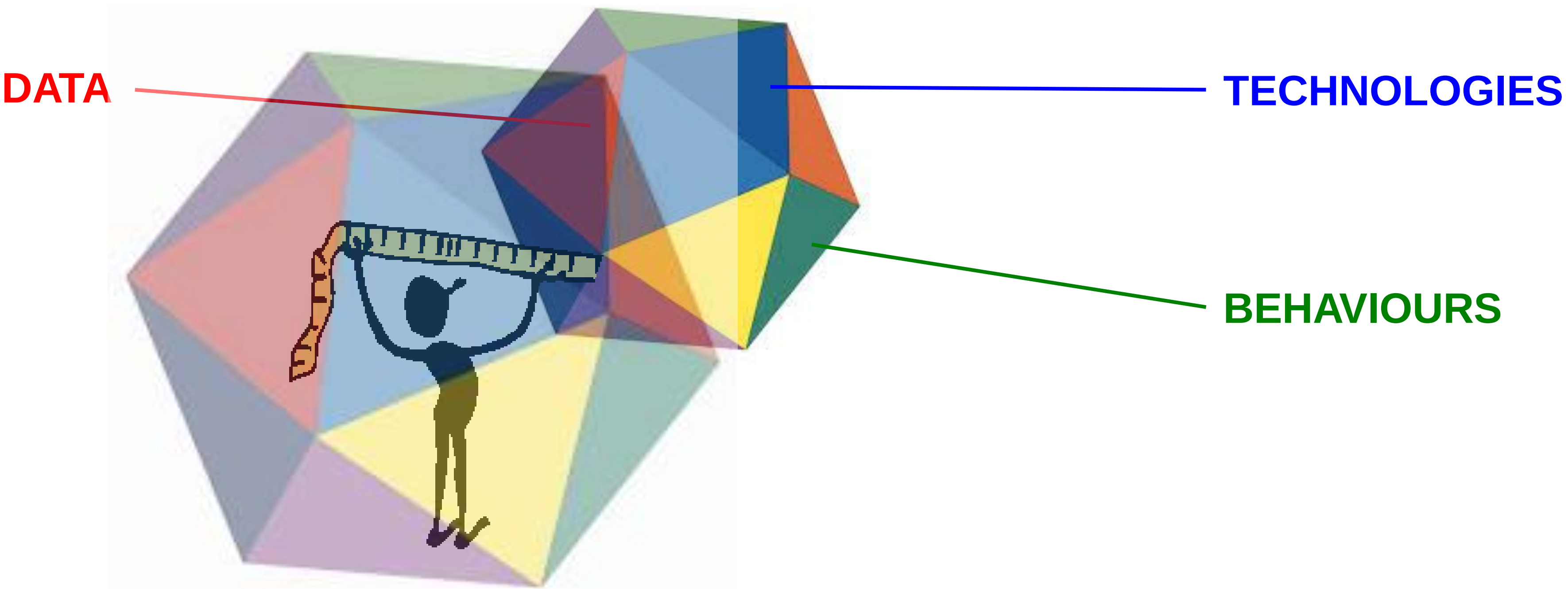
- What matters most is not that they are more/bigger, but that they are *different*(from traditional data sources)

  - Big data = non-traditional data = new digital data

Key point #2

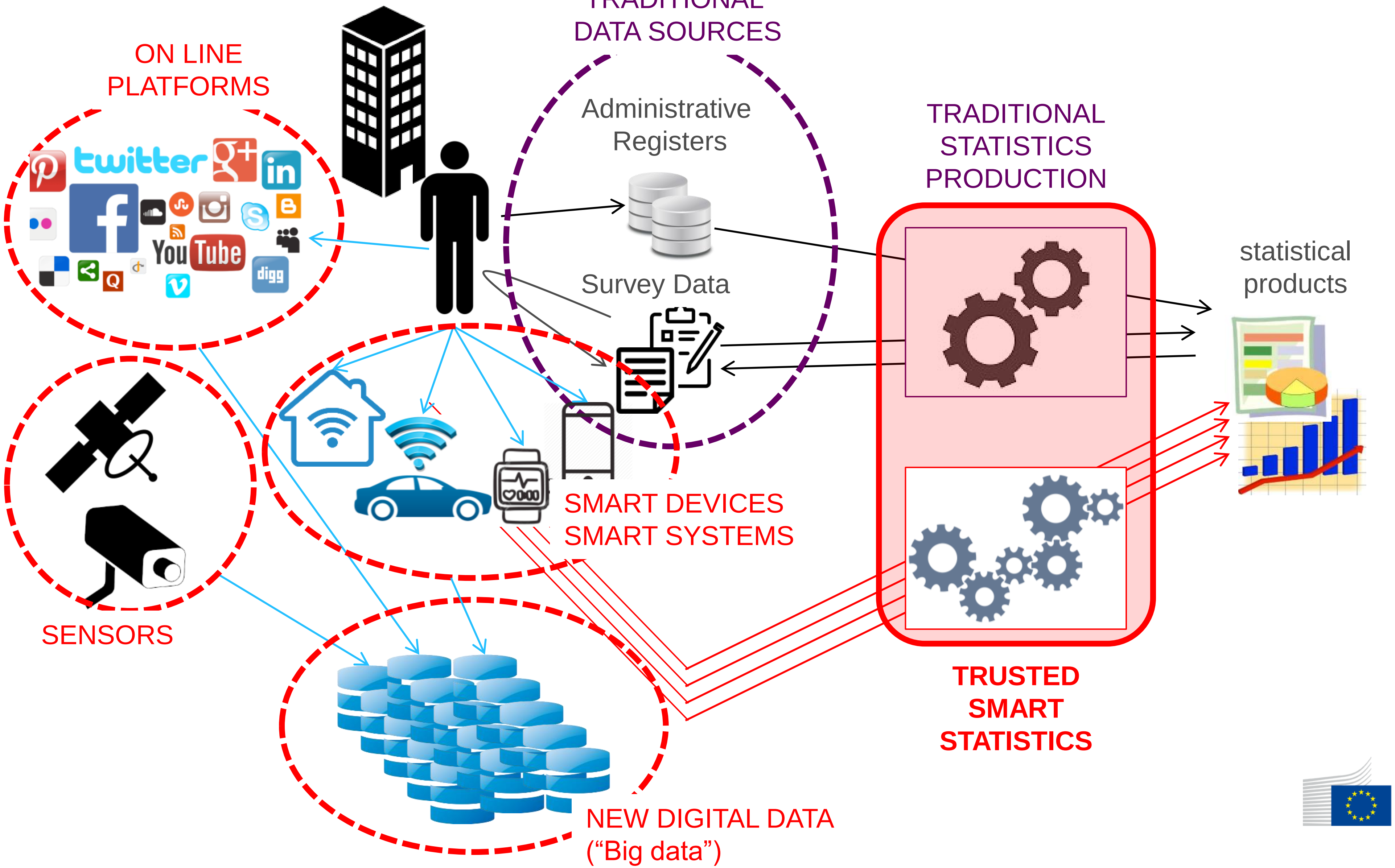
- New digital **data** come with new digital **technologies** and new digital **behaviours** and perceptions, attitudes, expectations …

- It's a new digital world - new data is one of its facets



Questions?

Trusted Smart Statistics
(the new engine)



TRADITIONAL DATA SOURCES

Administrative Registers

Survey Data

ON LINE PLATFORMS

SENSORS

SMART DEVICES
SMART SYSTEMS

NEW DIGITAL DATA
("Big data")

TRADITIONAL STATISTICS PRODUCTION

TRUSTED SMART STATISTICS

statistical products

Designing the new engine

- Trusted Smart Statistics (TSS)
  = systemic augmentation of official statistics

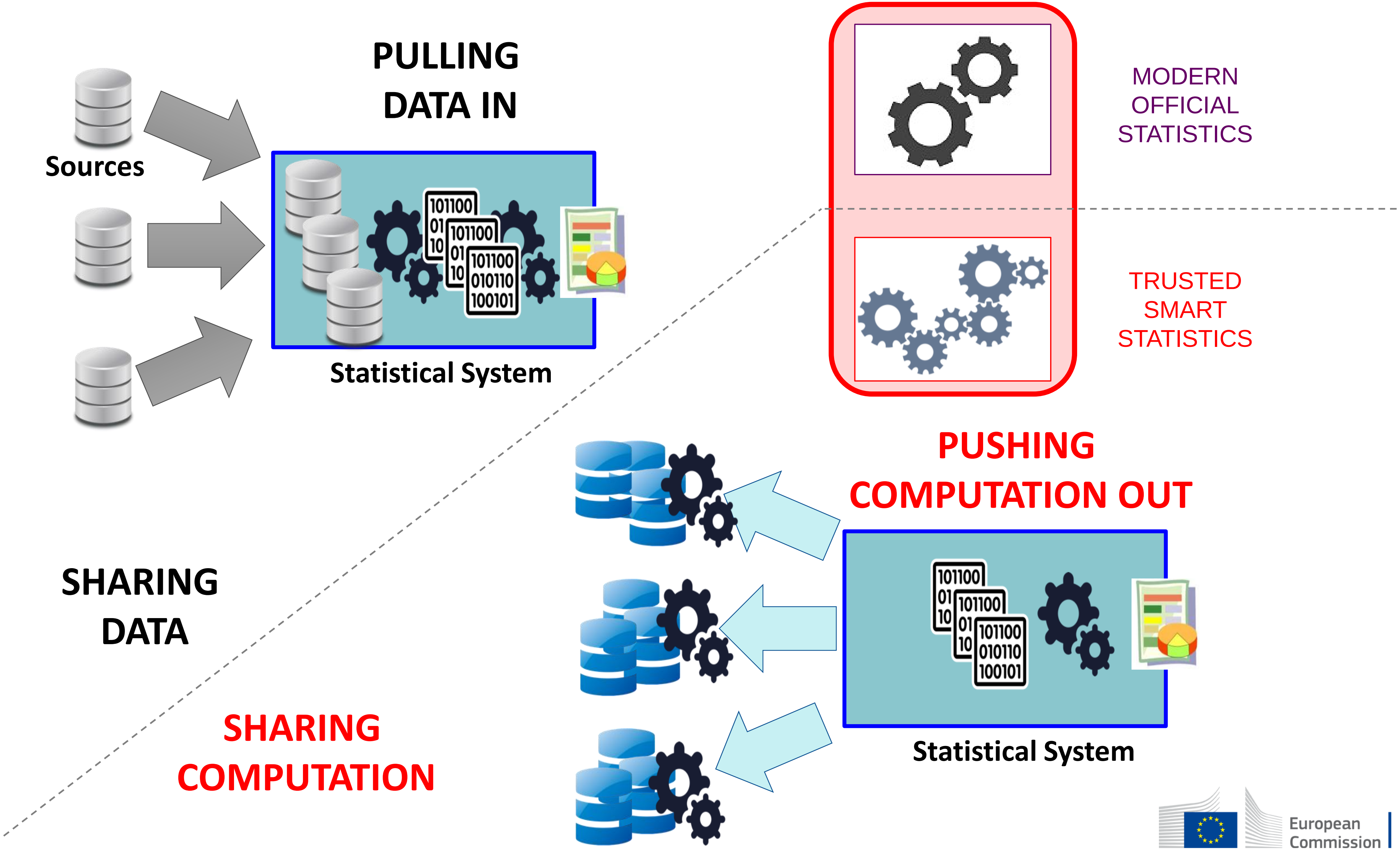
take a system-level view
define a clear “grand picture” first, then
develop components based on that…

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**

Design principles

1. Push computation out



PULLING DATA IN

Sources

Statistical System

MODERN OFFICIAL STATISTICS

TRUSTED SMART STATISTICS

PUSHING COMPUTATION OUT

Statistical System

SHARING DATA

SHARING COMPUTATION

Implications 1/2

Source code can be made publicly available, open-source
=> increase transparency, trustworthiness … and quality!

1. Push computation out

   - Requires **full automation**
     → methods encoded in machine-executable code
     (not just human-readable manuals)

   - Clear separation between methodological **development** (writing the source code)
     vs **production** (executing the binary code)

Methodological development always requires data exploration, hence “data in the house”.
But often can be performed on subsets of test data…

Implications 2/2

=> increase trustworthiness!

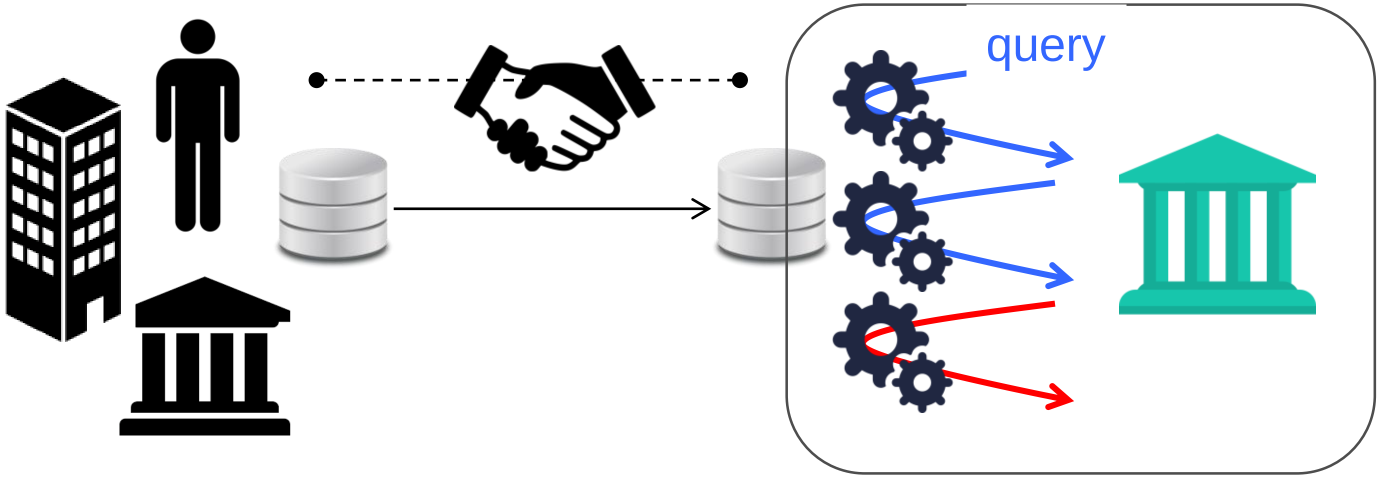
1. Push computation out

- sharing computation => sharing control
  (in the production phase)

- naturally combines with
  Secure Private Computing technologies
  (e.g. Secure Multi-Party Computation)

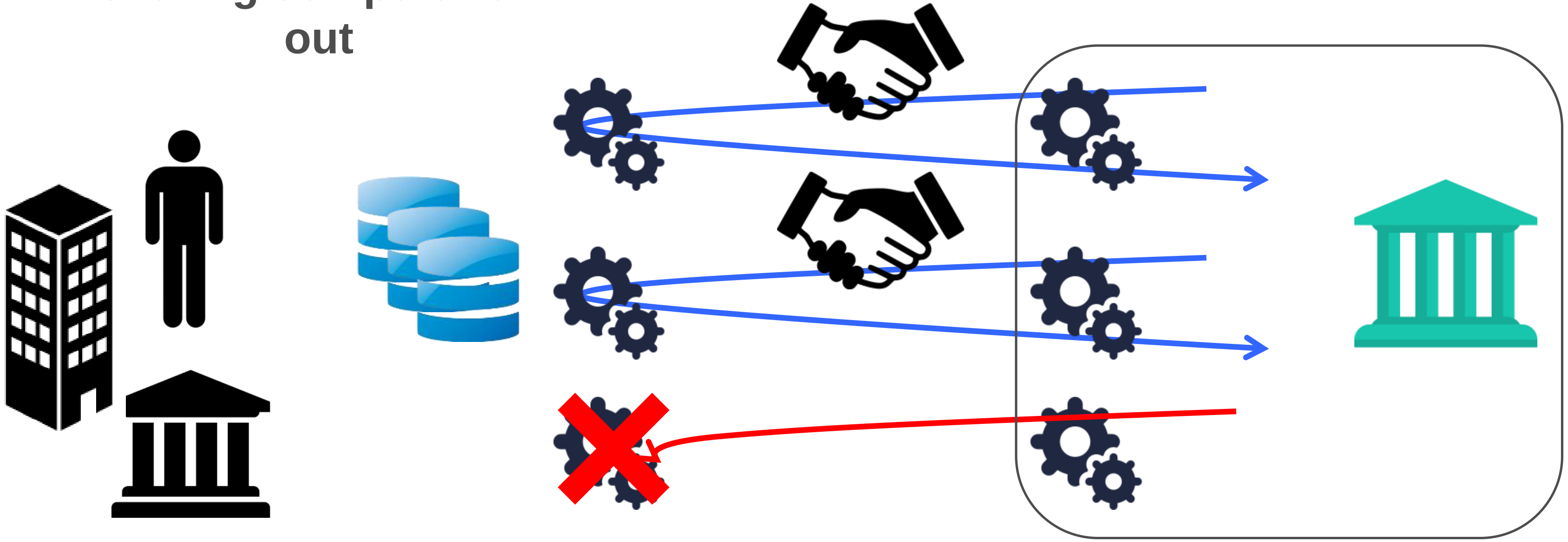
Increases protection of input data confidentiality
=> increase trustworthiness!

sharing computation => sharing control



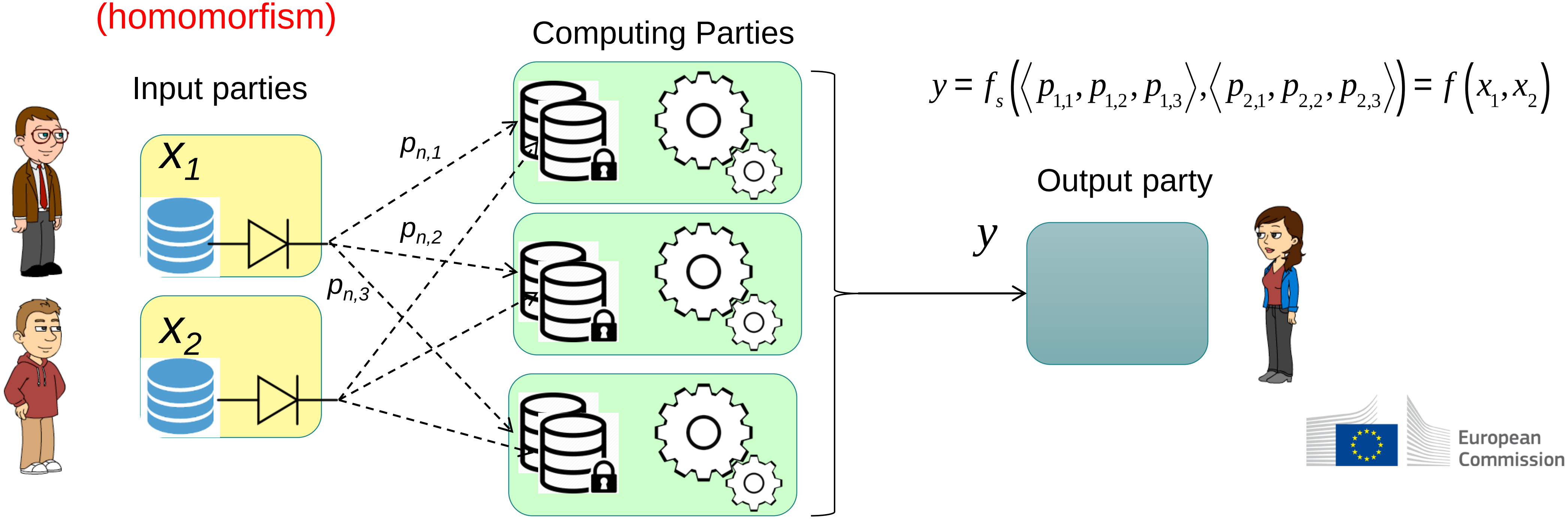
sharing data in

sharing computation out



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.
- 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)



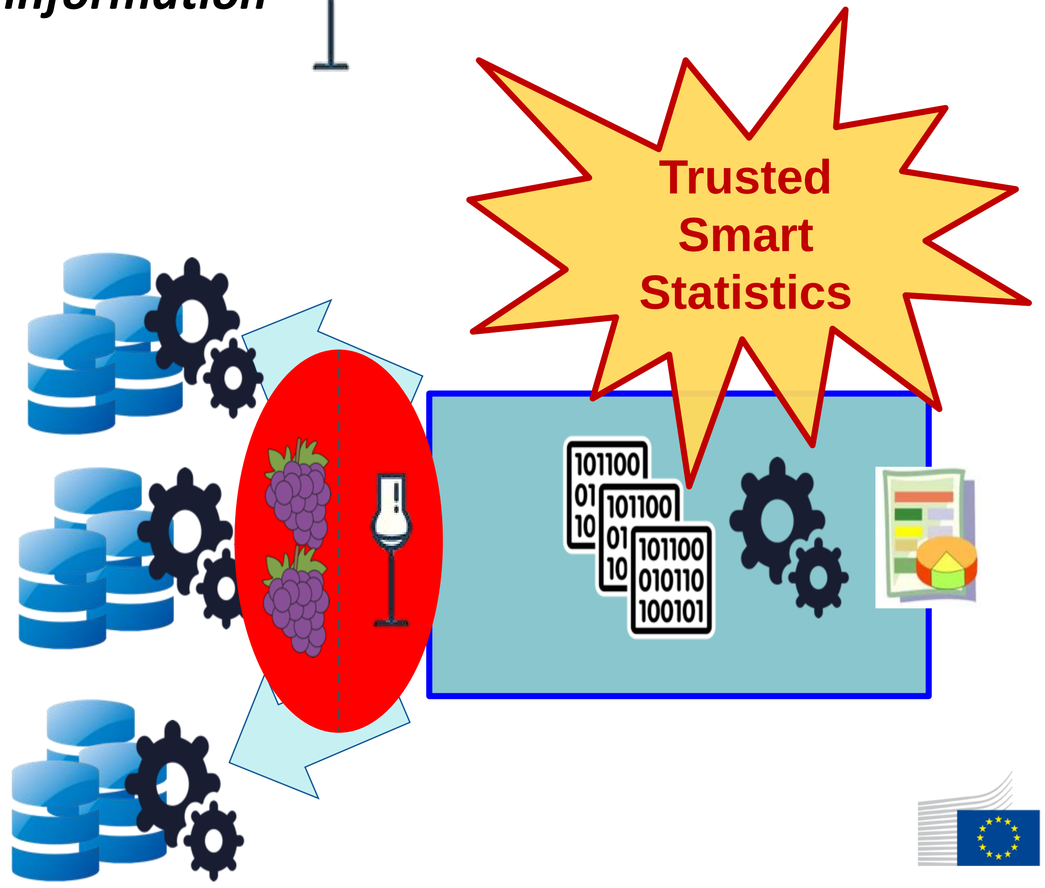
$$y = f_s\left(\left\langle p_{1,1}, p_{1,2}, p_{1,3}\right\rangle, \left\langle p_{2,1}, p_{2,2}, p_{2,3}\right\rangle\right) = f\left(x_1, x_2\right)$$

Secure Private Computing, transparency, auditability

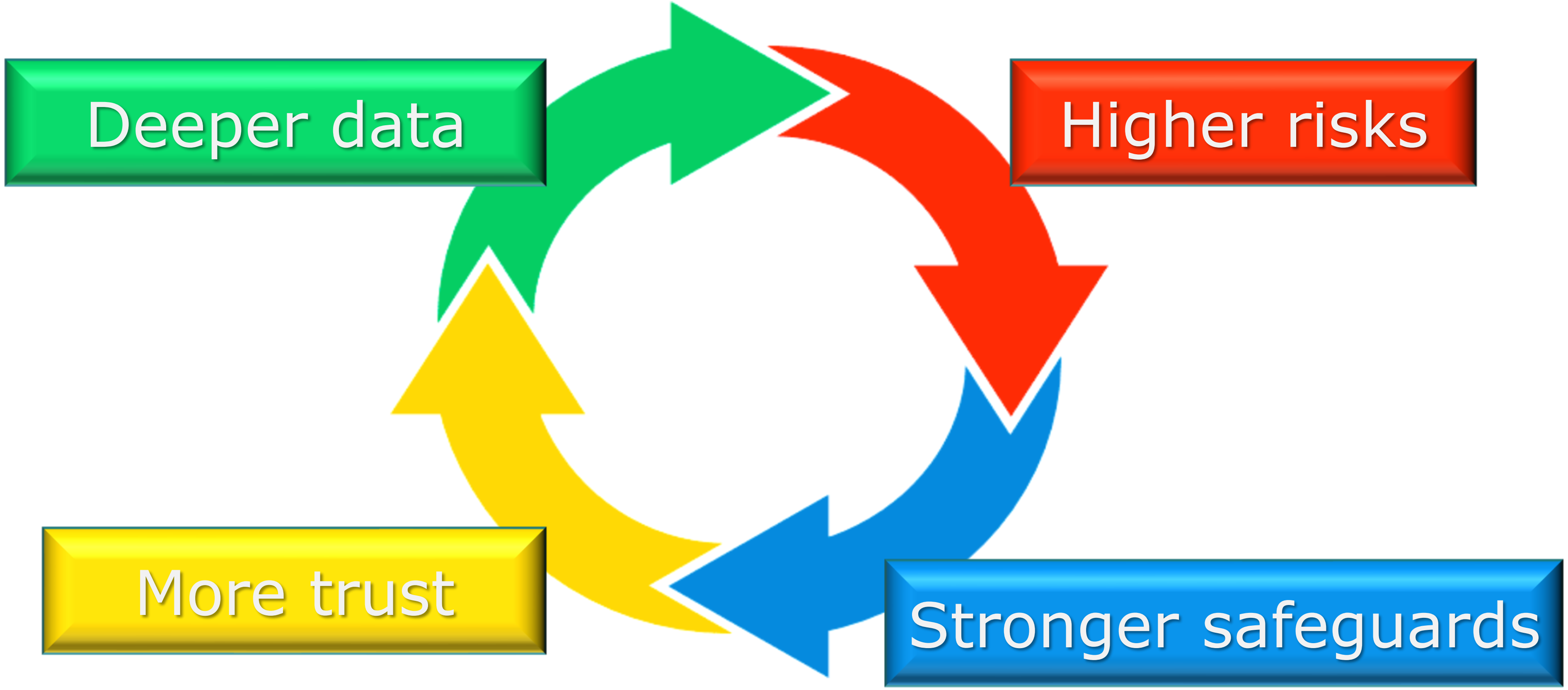
Adopt *Secure Private Computing* Technologies
(e.g. Secure Multi-Party Computation)
→ *disclosing only the desired* ***output information***
not the whole input data

Maximal transparency
→ *open-source code,*
non-modifiable logging of queries
→ *promote public scrutiny*

Sharing control with sources
over computation *execution*
→ *trust, participation, engagement*



To be SMART, you must be TRUSTED.
To be TRUSTED, you must be SMART.



Deeper data

Higher risks

Stronger safeguards

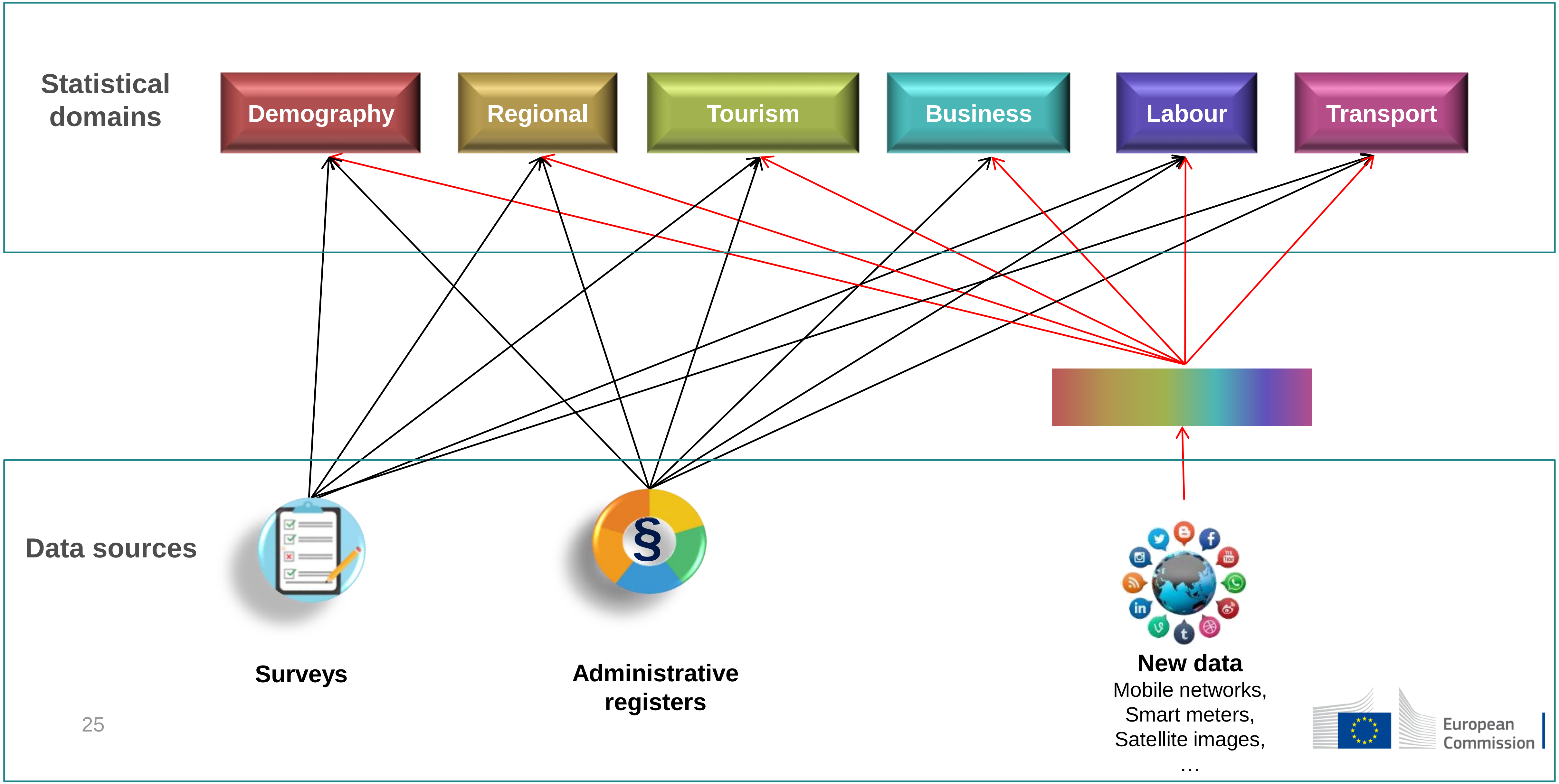
More trust

Questions?

Design Principles

1. Push computation out
2. Multi-purpose data sources for multi-source statistics

Multi-purpose data sources for multi-source statistics



Design Principles

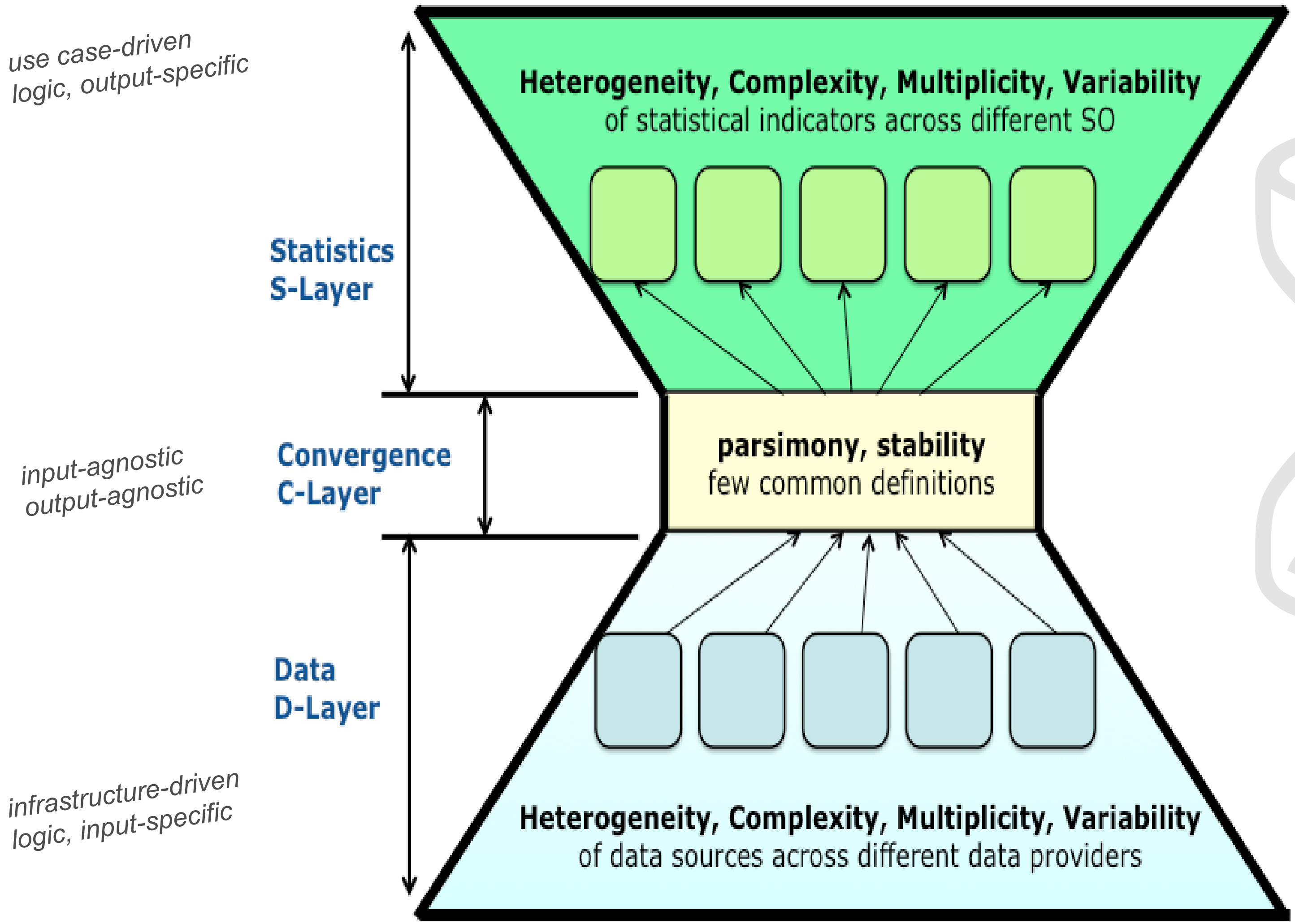
1. Push computation out
2. Multi-purpose data sources
   for multi-source statistics
3. Layered and modular organisation of the data workflow
   → Reference Methodological Frameworks

New business process, new functions

Multi-source statistical products

Multi-Purpose data

Layered approach, hourglass model



use case-driven logic, output-specific

Statistics S-Layer

Heterogeneity, Complexity, Multiplicity, Variability of statistical indicators across different SO

input-agnostic output-agnostic

Convergence C-Layer

parsimony, stability few common definitions

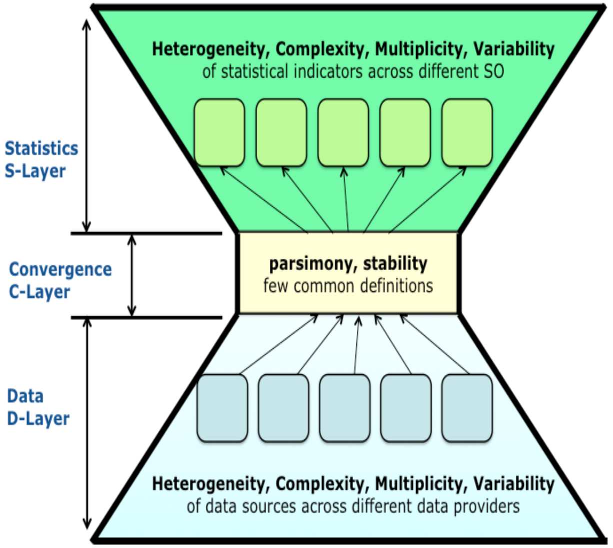
infrastructure-driven logic, input-specific

Data D-Layer

Heterogeneity, Complexity, Multiplicity, Variability of data sources across different data providers

Decoupling upper and bottom complexities

- **Complexity** of data semantic
  - domain-specific technological knowledge is required to extract the most/best information from raw data
- **Multiplicity & Heterogeneity**
  - different data providers
  - different data sources within each provider
  - different data formats, configurations
- **Variability**
  - data change following evolution of generating technology, infrastructure growth, reconfigurations, re-optimizations, SW releases …
  - socio-technological infrastructures are ever-evolving systems, not static objects



Heterogeneity, Complexity, Multiplicity, Variability
of statistical indicators across different SO

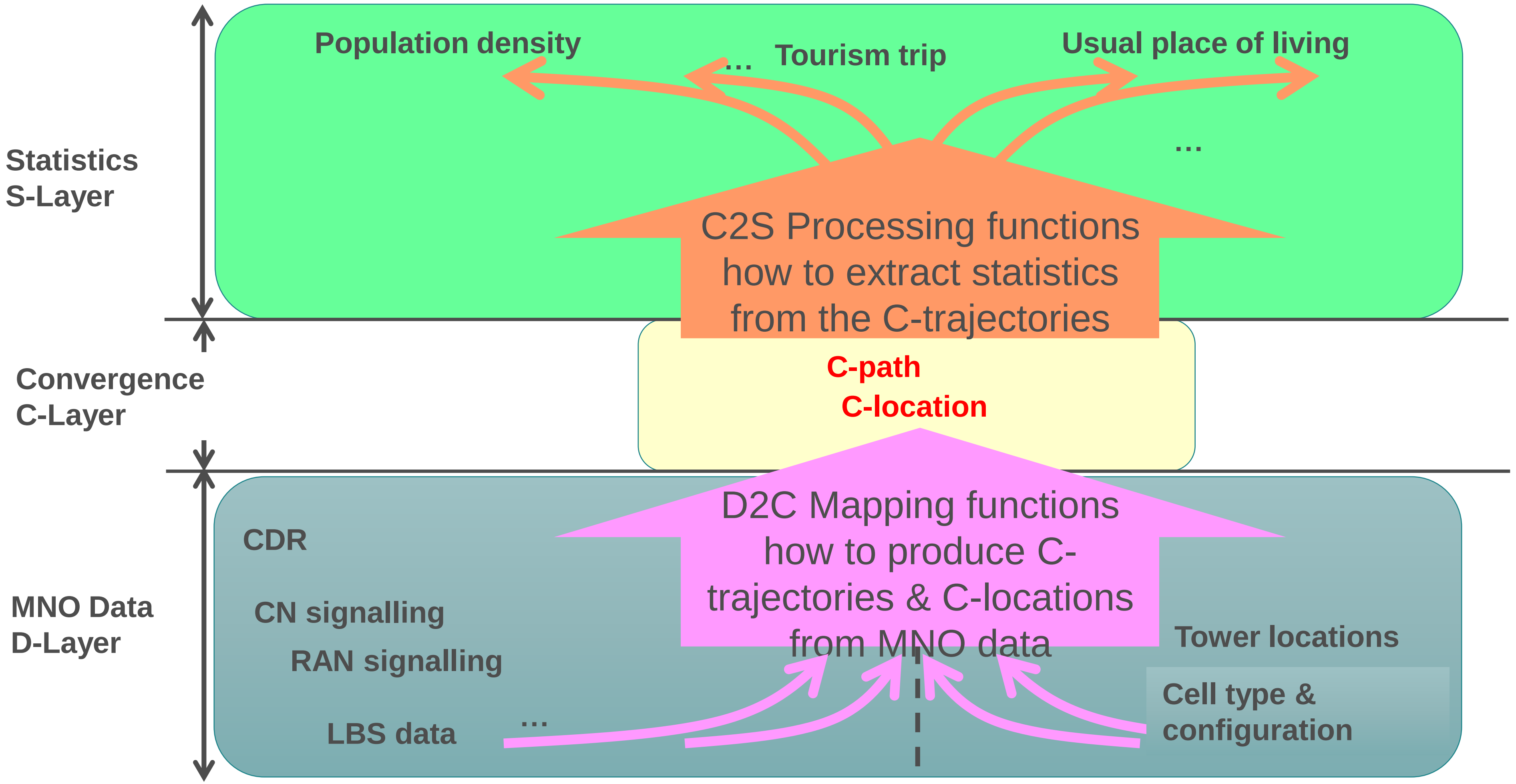
Statistics
S-Layer

parsimony, stability
few common definitions

Convergence
C-Layer

Data
D-Layer

Heterogeneity, Complexity, Multiplicity, Variability
of data sources across different data providers



Statistics
S-Layer
Convergence
C-Layer
MNO Data
D-Layer
Population density
… Tourism trip
Usual place of living
…
C2S Processing functions how to extract statistics from the C-trajectories
C-path
C-location
D2C Mapping functions how to produce C-trajectories & C-locations from MNO data
CDR
CN signalling
RAN signalling
LBS data
…
Tower locations
Cell type & configuration

European Commission

D2C Mapping functions

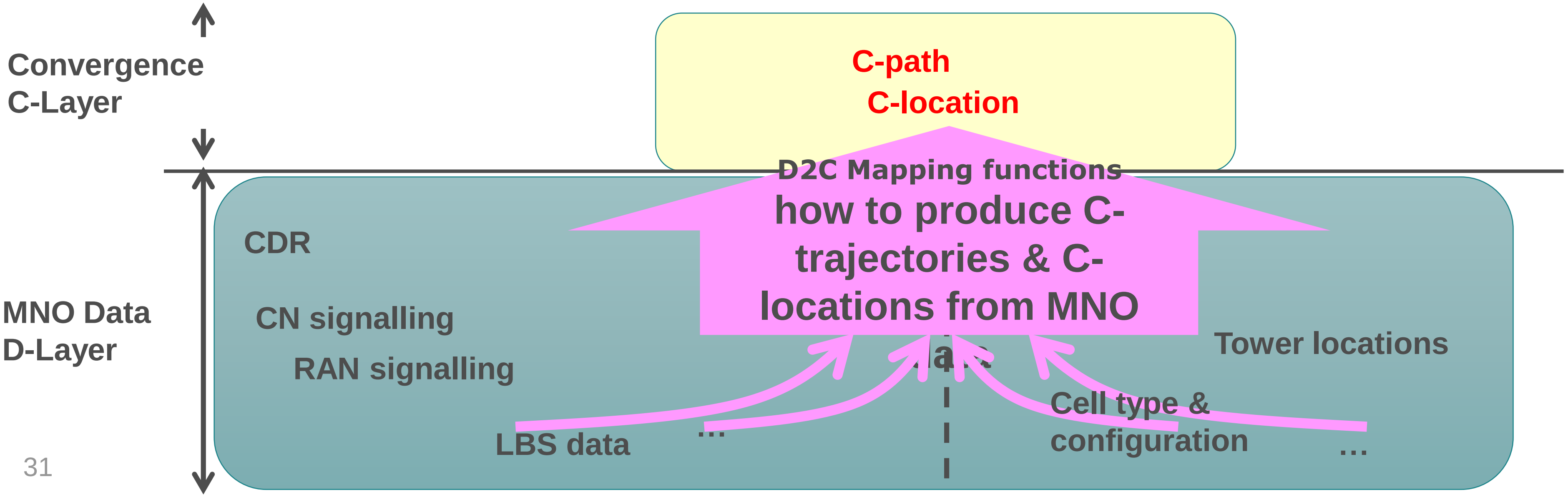
Technology-specific implementation of general principles.

Extract spatio-temporal information as accurately as possible given the available data.

Avoid distortion and/or loss of useful information.

Discard information not relevant for upper layers.

To be worked out by technology experts,
with support by statisticians



Convergence C-Layer

C-path
C-location

D2C Mapping functions
how to produce C-trajectories & C-locations from MNO

MNO Data D-Layer

CDR

CN signalling

RAN signalling

LBS data

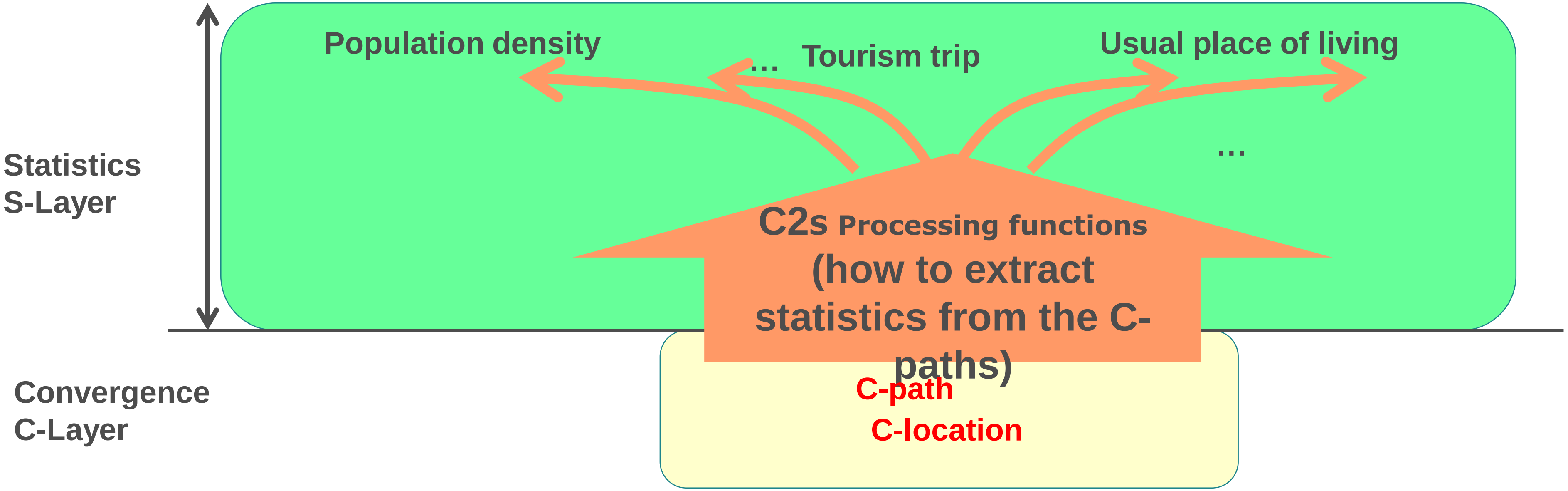
…

Tower locations

Cell type & configuration

…

C2S Processing functions



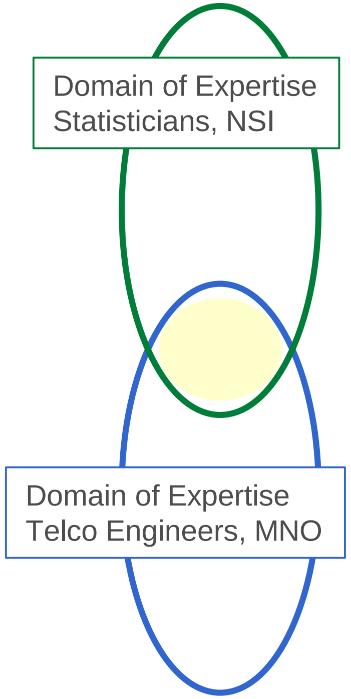
Statistical methods based on a sound understanding of C-layer data and meta-data (semantic, sources of errors).

To be worked out by statisticians,

with support by technology experts

C-layer structures

- i.e., **data** with a “normalized” semantic (future-proof, statistician-friendly,…)
  - … based on a parsimonious data generation model that includes (implicitly or explictly) the relevant sources of error, uncertainty, limitations to resolution, etc.

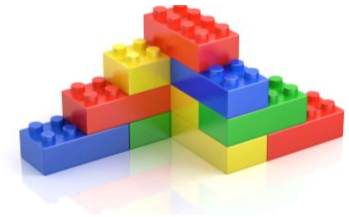


Domain of Expertise
Statisticians, NSI

Domain of Expertise
Telco Engineers, MNO

- and **meta-data**
  - … including quantitative indicators of error levels, resolution, uncertainty, etc.

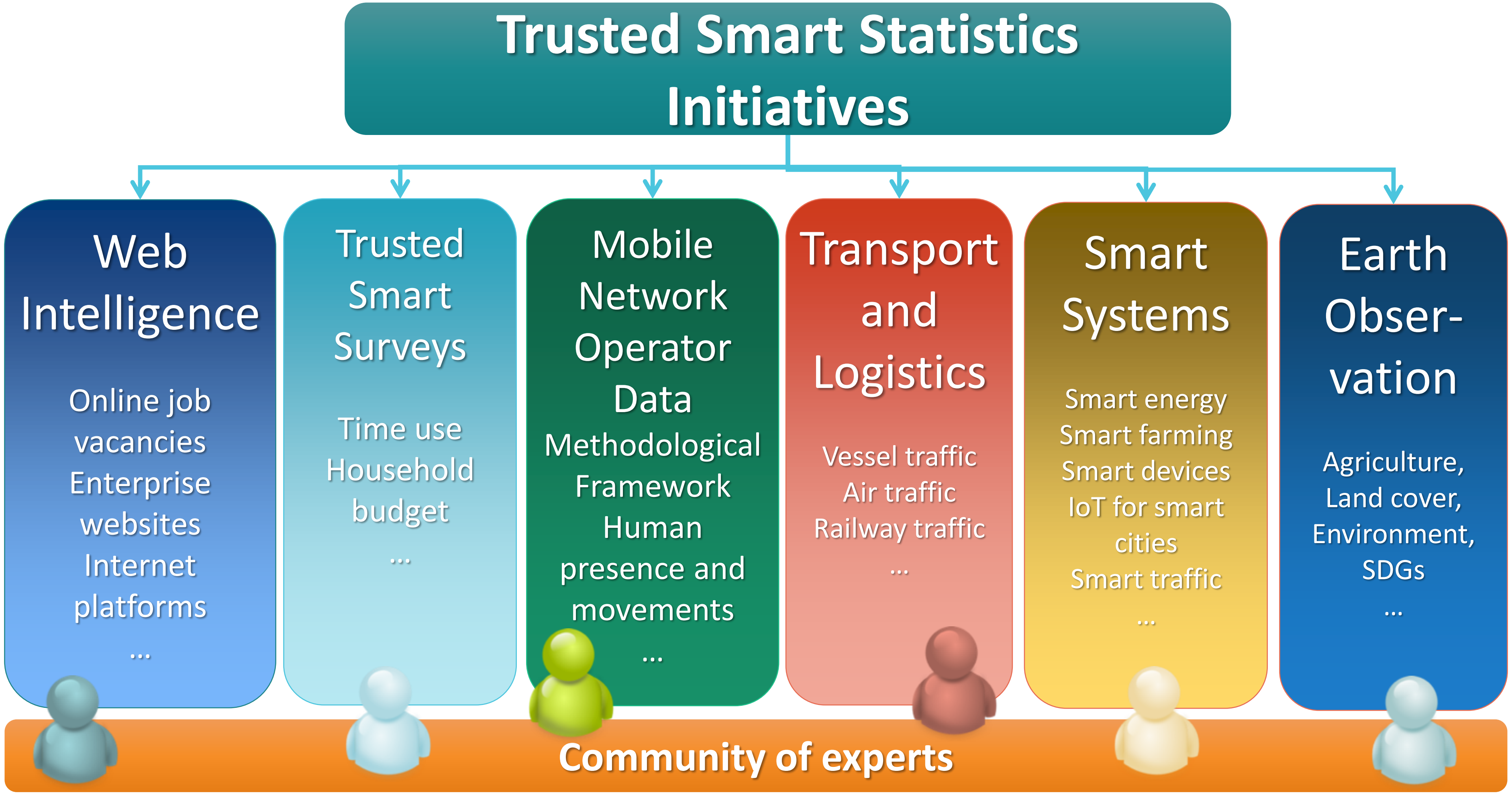
Take home message



- **The new fuel needs a new engine**
  - Exploiting “new (big) data” for Official Statistics requires a **new paradigm**: Trusted Smart Statistics
- System-level view: hardware, software, humanware
  - New technological solutions to ensure data confidentiality and process transparency
  - Import best practices from other fields: open-source algorithms, engagement with prod-users, citizen science -> citizen statistics
- Methodological work is needed
  - New (modular) reference methodological frameworks for new data sources
  - Design for evolvability – of algorithms and data
  - Co-development by statisticians and technology experts needed
  - Using new data sources requires investments in methodology (and infrastructure)

Questions?

Eurostat initiatives on Trusted Smart Statistics



Trusted Smart Statistics Initiatives

Web Intelligence
Online job vacancies
Enterprise websites
Internet platforms
…

Trusted Smart Surveys
Time use
Household budget
…

Mobile Network Operator Data
Methodological Framework
Human presence and movements
…

Transport and Logistics
Vessel traffic
Air traffic
Railway traffic
…

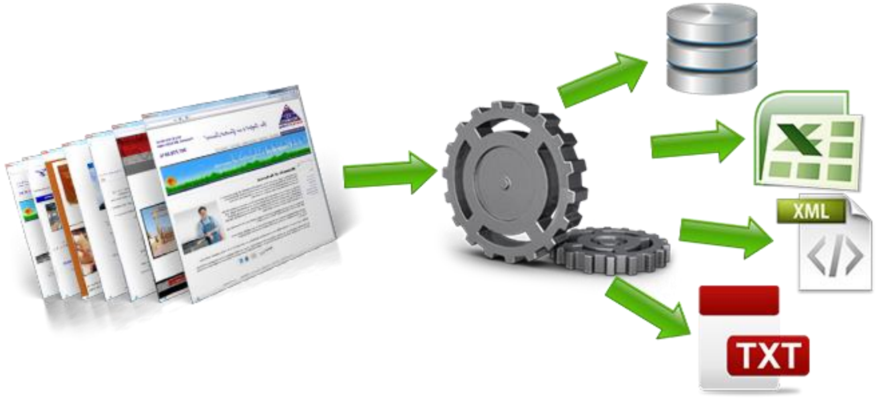
Smart Systems
Smart energy
Smart farming
Smart devices
IoT for smart cities
Smart traffic
…

Earth Observation
Agriculture, Land cover, Environment, SDGs
…

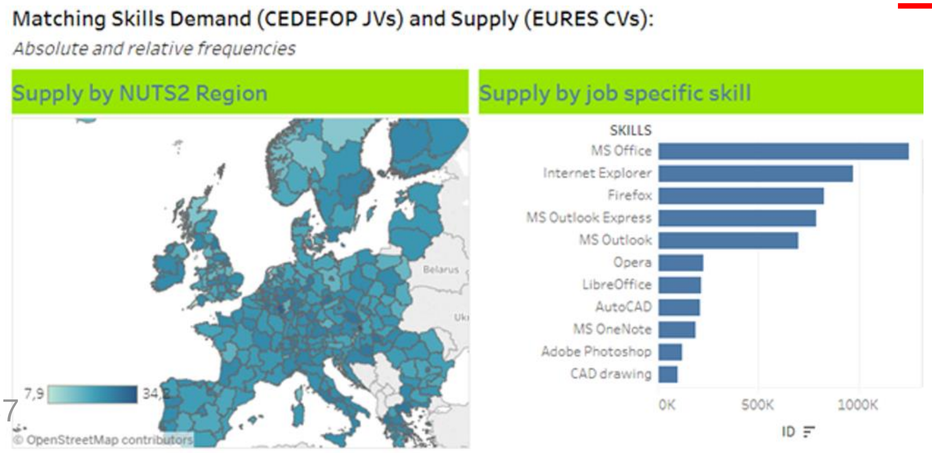
Community of experts

Web Intelligence Hub

✓ A bundle of capabilities to support the collection, processing, reuse and analysis of web data ressource (web pages, APIs …) for producing statistics



- ○ Online job vacancies advertisement
  - •Skills, job vacancies
- ○ Enterprise websites
  - •Business registers, jobs, information society


Matching Skills Demand (CEDEFOP JVs) and Supply (EURES CVs):

Absolute and relative frequencies

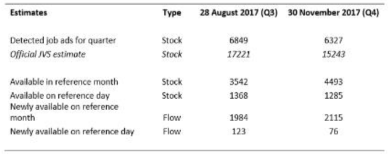
Supply by NUTS2 Region

Supply by job specific skill

Web Intelligence Hub – Expected benefits

- ➢ Complementary statistical products
- ➢ Improved statistical outputs
- ➢ Increased spatial granularity
- ➢ Flexible and interactive dashboarding
- ➢ Shared solutions
- ➢ …

Table 3: Experimental on-line job vacancy statistics for Slovenia



Estimates	Type	28 August 2017 (Q3)	30 November 2017 (Q4)
Detected job ads for quarter	Stock	6849	6327
Official JVS estimate	*Stock*	*17221*	*15243*
Available in reference month	Stock	3542	4493
Available on reference day	Stock	1368	1285
Newly available on reference month	Flow	1984	2115
Newly available on reference day	Flow	123	76

Figure 5: Nowcasts based on the S-ARIMA-X time series model.

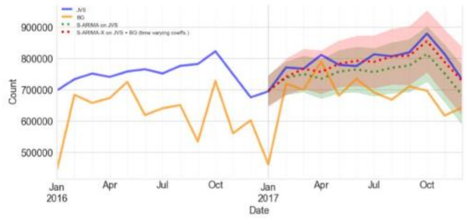


Figure 4a: Time series of the total JV counts, averaged per month (scaled to the JVS scale).

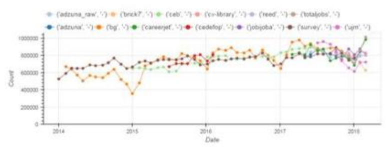


Figure 4b: Time series of the JVS and daily average of the online sources (scaled to the JVS)

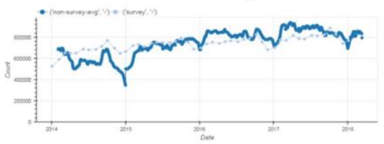
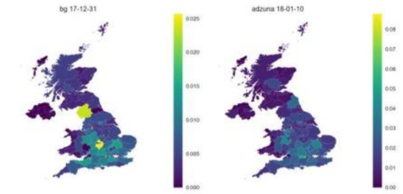
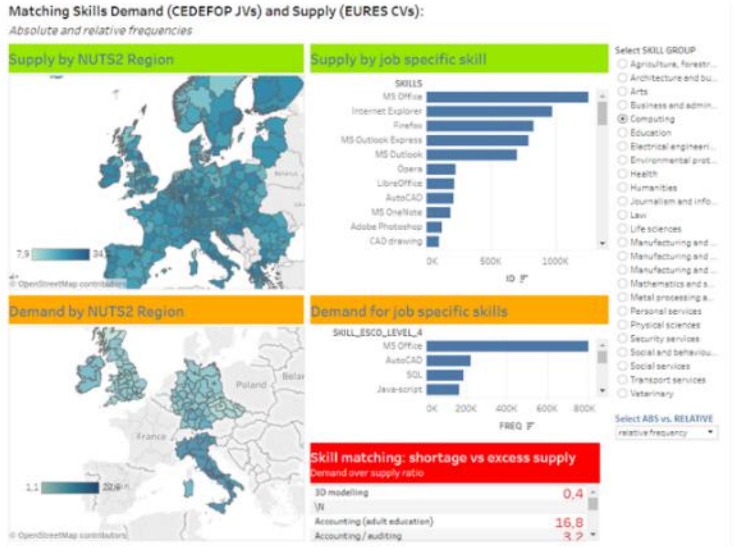


Figure 6: Number of job vacancies as a proportion of working age population[28]





Matching Skills Demand (CEDEFOP JVs) and Supply (EURES CVs):

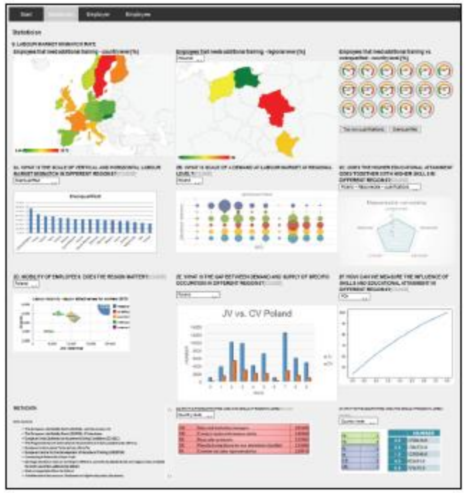
Absolute and relative frequencies

Select SKILL GROUP
○ Agriculture, forestr...
○ Architecture and bu...
○ Arts
○ Business and admin...
● Computing
○ Education
○ Electrical engineeri...
○ Environmental prot...
○ Health
○ Humanities
○ Journalism and info...
○ Law
○ Life sciences
○ Manufacturing and ...
○ Manufacturing and ...
○ Manufacturing and ...
○ Mathematics and s...
○ Metal processing a...
○ Personal services
○ Physical sciences
○ Security services
○ Social and behaviou...
○ Social services
○ Transport services
○ Veterinary

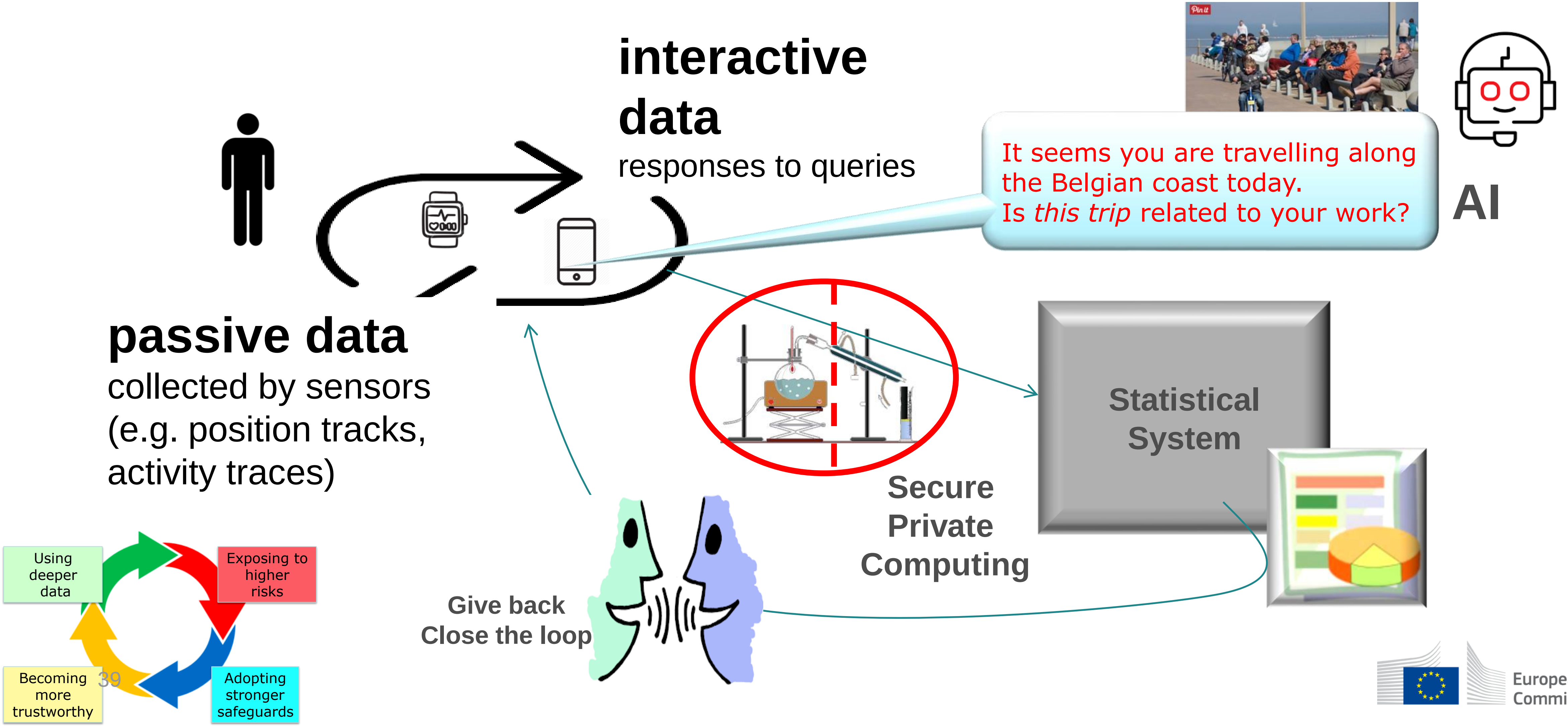
Select ABS vs. RELATIVE: relative frequency

Skill matching: shortage vs excess supply
Demand over supply ratio

3D modelling	0,4
\N	
Accounting (adult education)	16,8
Accounting / auditing	3,2



Trusted Smart Surveys, Citizen Statistics



interactive data

responses to queries

It seems you are travelling along the Belgian coast today.
Is *this trip* related to your work?

AI

passive data

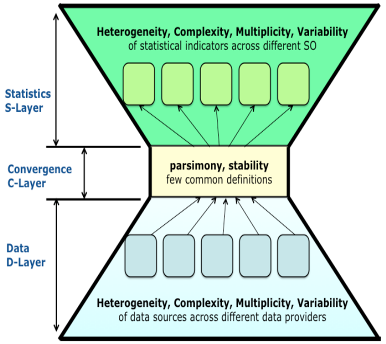
collected by sensors (e.g. position tracks, activity traces)

Secure Private Computing

Statistical System

Give back
Close the loop

Mobile Network Operator (MNO) Data



Statistics
S-Layer

Heterogeneity, Complexity, Multiplicity, Variability
of statistical indicators across different SO

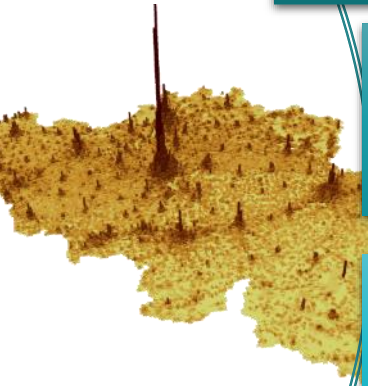
Convergence
C-Layer

parsimony, stability
few common definitions

Data
D-Layer

Heterogeneity, Complexity, Multiplicity, Variability
of data sources across different data providers

Develop a methodological framework and robust methodologies for selected use-cases



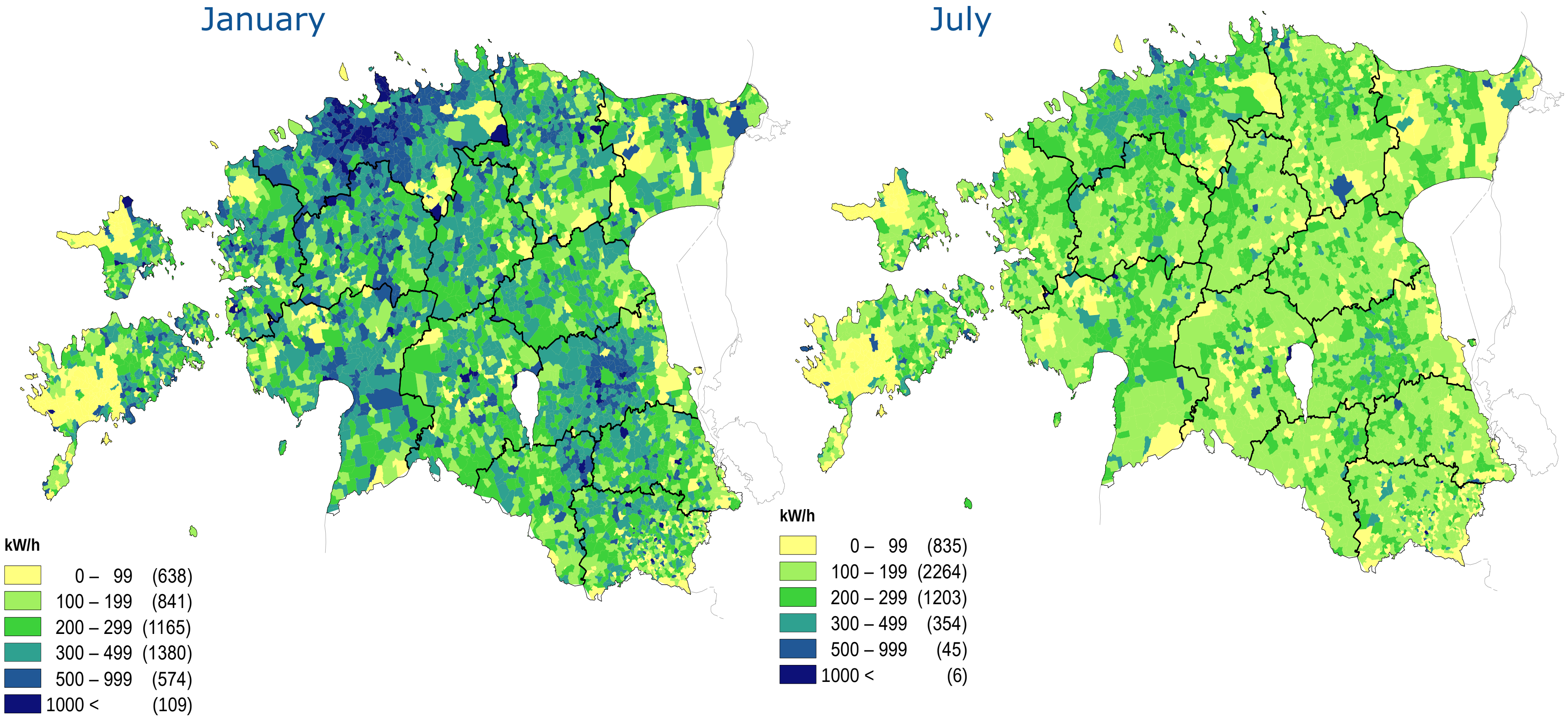
Build expert knowledge about mobile network technologies.

Pilot applications of Privacy-Enhancing Technologies
Pilot multi-MNO deployments

Initial focus on population and tourism statistics

Smart Systems: Electricity Meters

Household Consumption per Commune

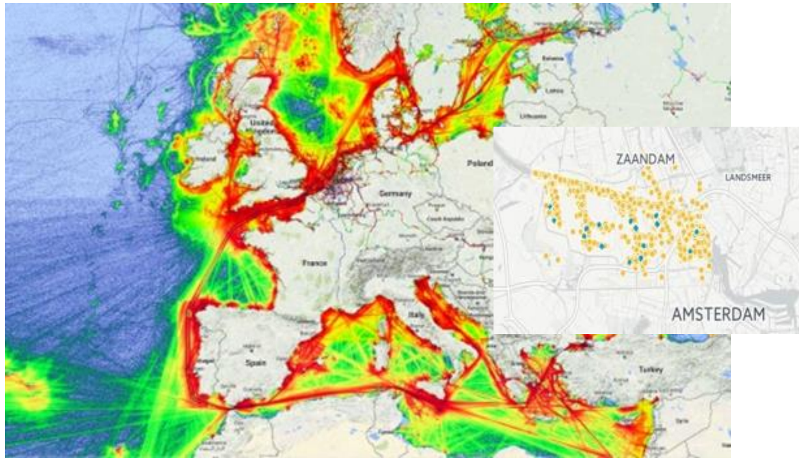


January

July

Transport and logistics

Use of tracking data to provide long-distance transportation and logistics.

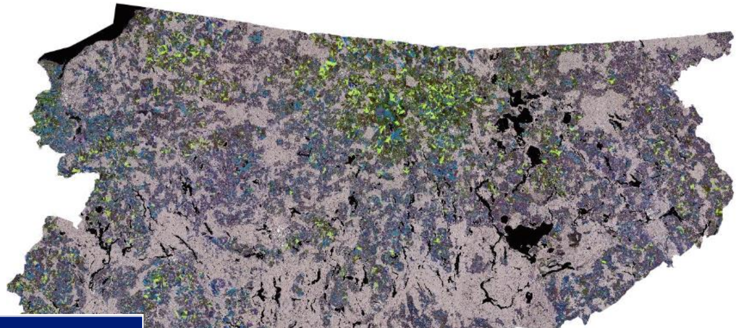


Initial focus on
Ship position data

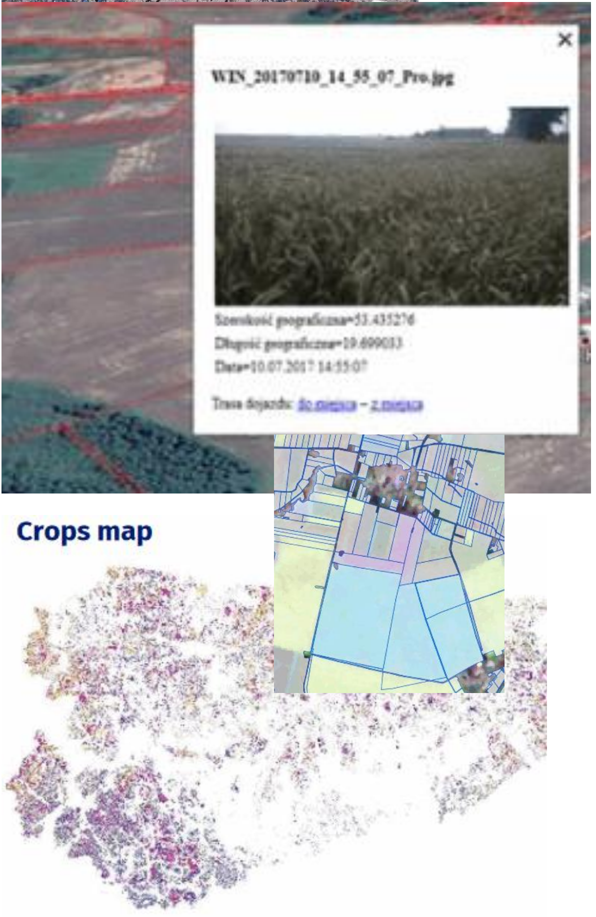
Extension to air and railway traffic data

Flash estimates of economic indicators

Earth Observation



Agriculture	Case study 1	*Crop recognition, mapping and monitoring*
	Case study 2	*Monitoring of the off-season vegetation cover*
	Case study 3	*Crop recognition with very high resolution aerial data*
Build-up area	Case study 4	*Implementing SDG indicator 11.7.1*
	Case study 5	*Urban sprawl across urban areas in Europe*
	Case study 6	*Combination of administrative and Earth Observation data to determine the quality of housing*
Land cover	Case study 7	Comparing «in-situ» and «remote-sensing» collection mode for land cover data
	Case study 8	*Land cover maps at very detailed scale*
Settlements, Enumeration Areas and Forestry	Case study 9	*Update the INSPIRE Theme Statistical Units dataset and preventing forest fire*



WIN_20170710_14_55_07_Pro.jpg

Szerokość geograficzna=53.435276

Długość geograficzna=19.699033

Data=10.07.2017 14:55:07

Trasa dojazdu: do miejsca – z miejsca

European Commission

Statistical Journal of the IAOS

Journal of the International Association for Official Statistics

WORLDOFSTATISTICS.ORG THE WORLD OF STATISTICS

IAOS

Special section on ‘Trusted Smart Statistics’

A special section in this issue of the Journal is dedicated to nine manuscripts on the very current topic of ‘**Trusted Smart Statistics**’. This section gathers extended versions of papers that were presented at the 104th DGINS conference in October 2018, held in Bucharest (Rumania). The section illustrates how the European Statistical System (ESS) calls the future of Official Statistics and how in operational terms the concern for maintaining and improving trust is included in the production and dissemination of statistics. The section is introduced in **a guest editorial by Mariana Kotzeva**, the Director General of Eurostat.

The first section on **'The future role of Official Statistics in the informational ecosystem'** which is the leading topic for this issue and also the item for the second discussion on the discussion platform (officialstatistics.com/discussion-platform) has been discussed in the first newsletter on this issue. The fourth section of this issue will be highlighted in the next newsletter (March).

Further Reading

- F. Ricciato, A. Wirthmann, K. Giannakouris, F. Reis, and M. Skaliotis. Trusted smart statistics: Motivations and principles. Statistical Journal of the IAOS, 35(4), 2019. https://ec.europa.eu/eurostat/cros/system/files/sji190584.pdf

- F. Ricciato, G. Lanzieri, A. Wirthmann, G. Seynaeve. Towards a methodological framework for estimating present population density from mobile network operator data, working paper, an earlier version was presented to the IUSSP workshop on digital demography, Seville, June 2019, https://ec.europa.eu/eurostat/cros/system/files/mno_spatial_density_ricciato_lanzieri_wirthmann_2020_v2.pdf

- F. Ricciato. Towards a reference methodological framework for processing MNO data for official statistics. In15th Global Forum on Tourism Statistics, Cusco, Peru, November 2018. https://tinyurl.com/ycgvx4m6

- F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano. Estimating population density distribution from network-based mobile phone data. JRCTechnical Report, 2015. https://tinyurl.com/ydz4mgaw

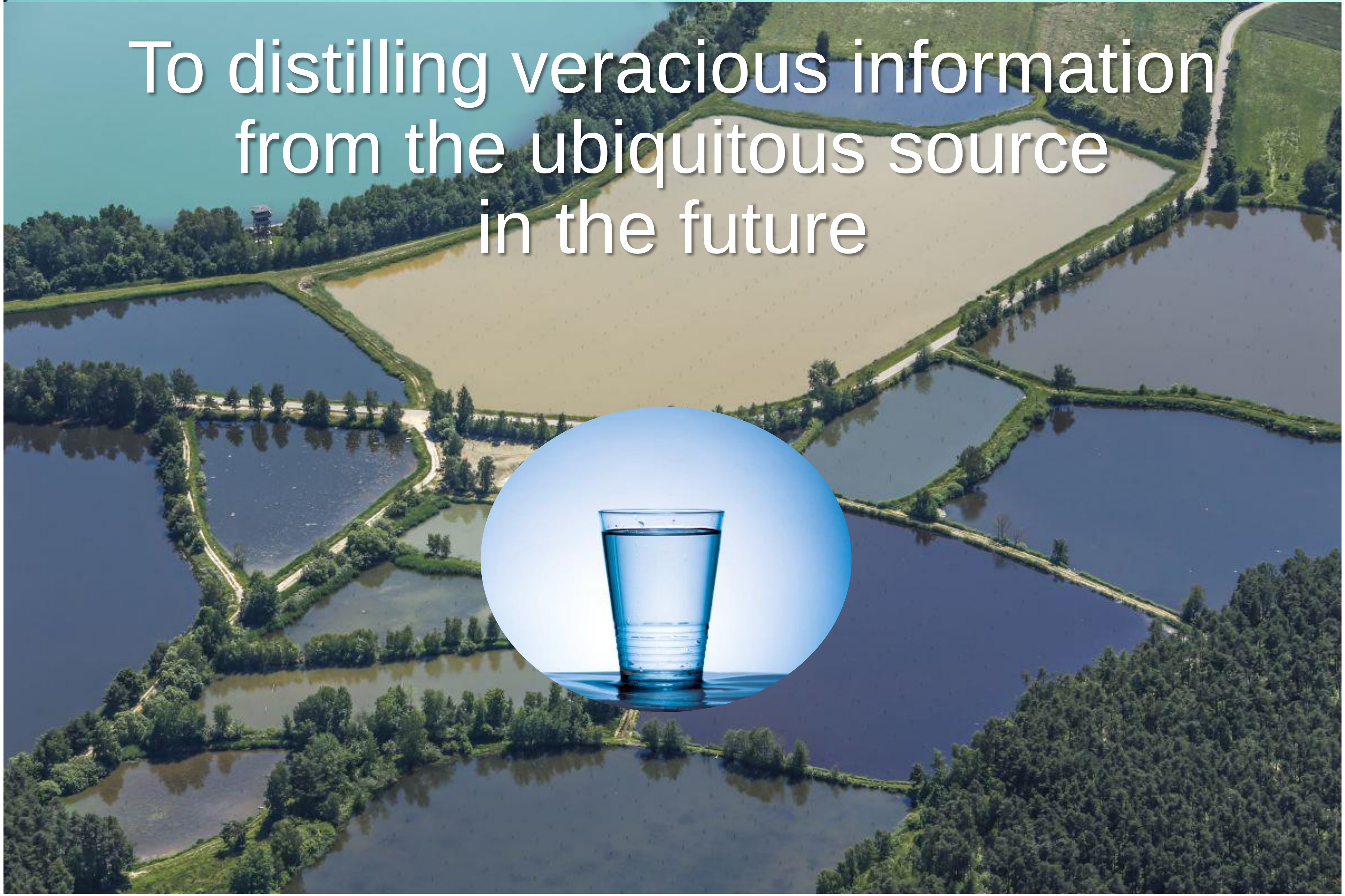
- Big Data UN Global Working Group. Un handbook on privacy-preservingcomputation techniques https://tinyurl.com/y3rg5azm, 2019.

European Commission

Data: a scarce commodity in the past

From concentrating efforts on collecting data,

To distilling veracious information from the ubiquitous source in the future



Thank you

Slide xx: element concerned, source: e.g. Fotolia.com; Slide xx: element concerned, source: e.g. iStock.com