Workshop on Methodologies for Official Statistics

5 - 6 December 2022, Rome (Istat)

Methodologies for big data
Overview of Istat's activities and open challenges

Mauro Bruno

Directorate for Methodology and Statistical Process Design (DCME)

Italian National Institute of Statistics – Istat

Monica Scannapieco

Italian National Institute of Statistics – Istat (currently working at AGENZIA PER LA CYBERSICUREZZA NAZIONALE)

Outline

- Big Data @Istat: Why, When, How
- Big Data Projects @Istat:
  - **Text Processing** Pipelines
  - **Image Processing** Pipelines
  - Improving data dissemination **timeliness**
- Conclusions

Big Data @Istat:
Why, When, How

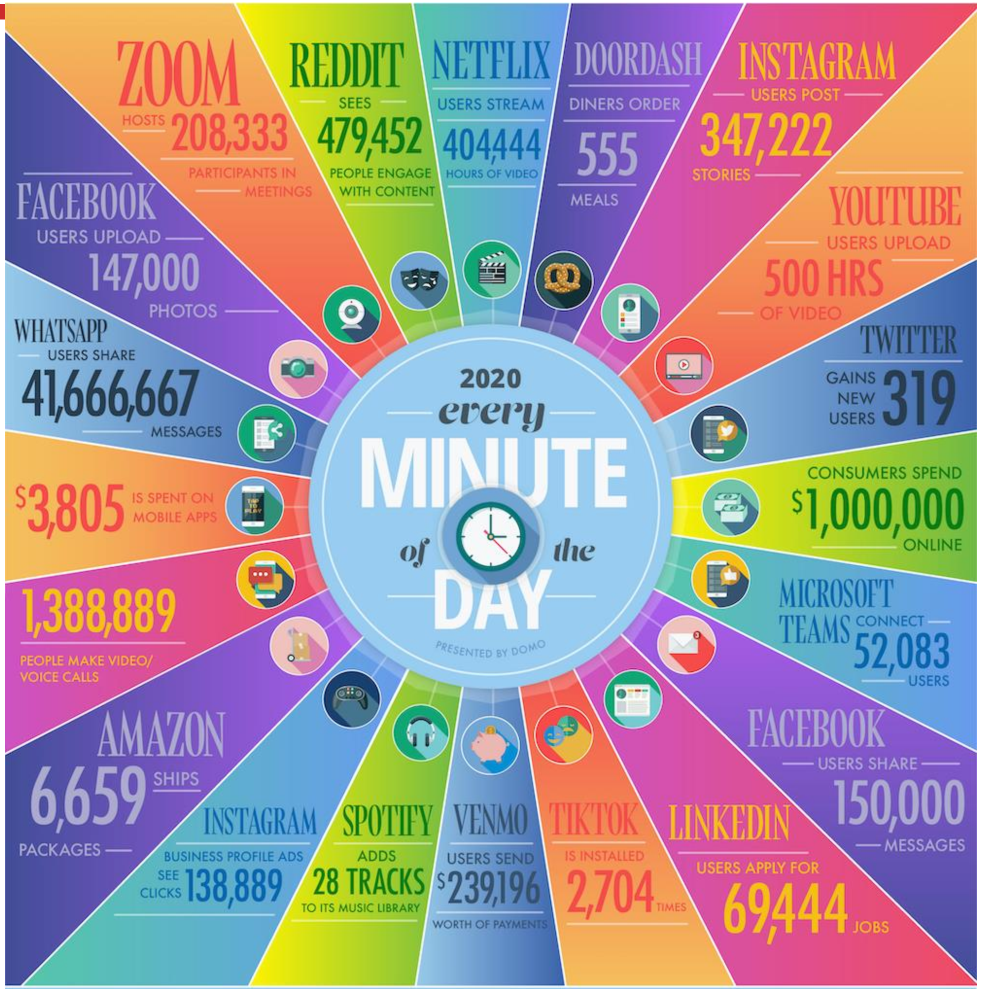
Big Data @Istat: Why & When

○ European Statistical System strategic drivers

- Scheveningen Memorandum “Big Data in Official Statistics” - 2013
- Bucharest Memorandum “**Official Statistics in a datafied society - Trusted Smart Statistics**” - 2018

○ Main Objectives

- Enrich statistical production with **new products**
- **Enhance timeliness** in official statistical production
- Official statistics relevance in **new data ecosystem**



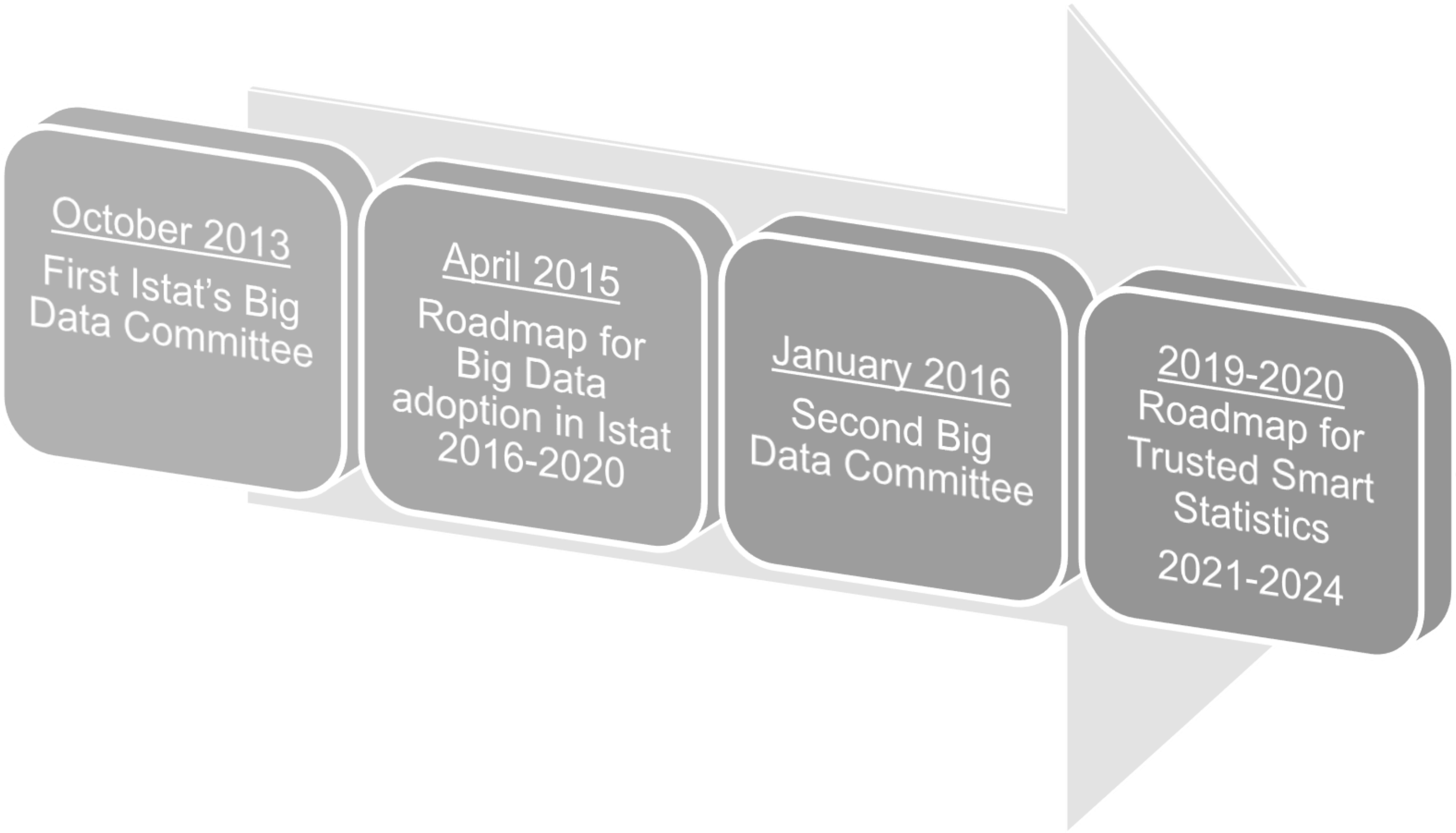
Source: Data Never Sleeps 8.0

Big Data @Istat: How

○ Istat's strategic context:

The use of Big Data in Official Statistics requires **methodological**, **technological** and **organizational** investments.

Starting from 2013, Istat created a Big Data Committee responsible of the Big Data strategy…

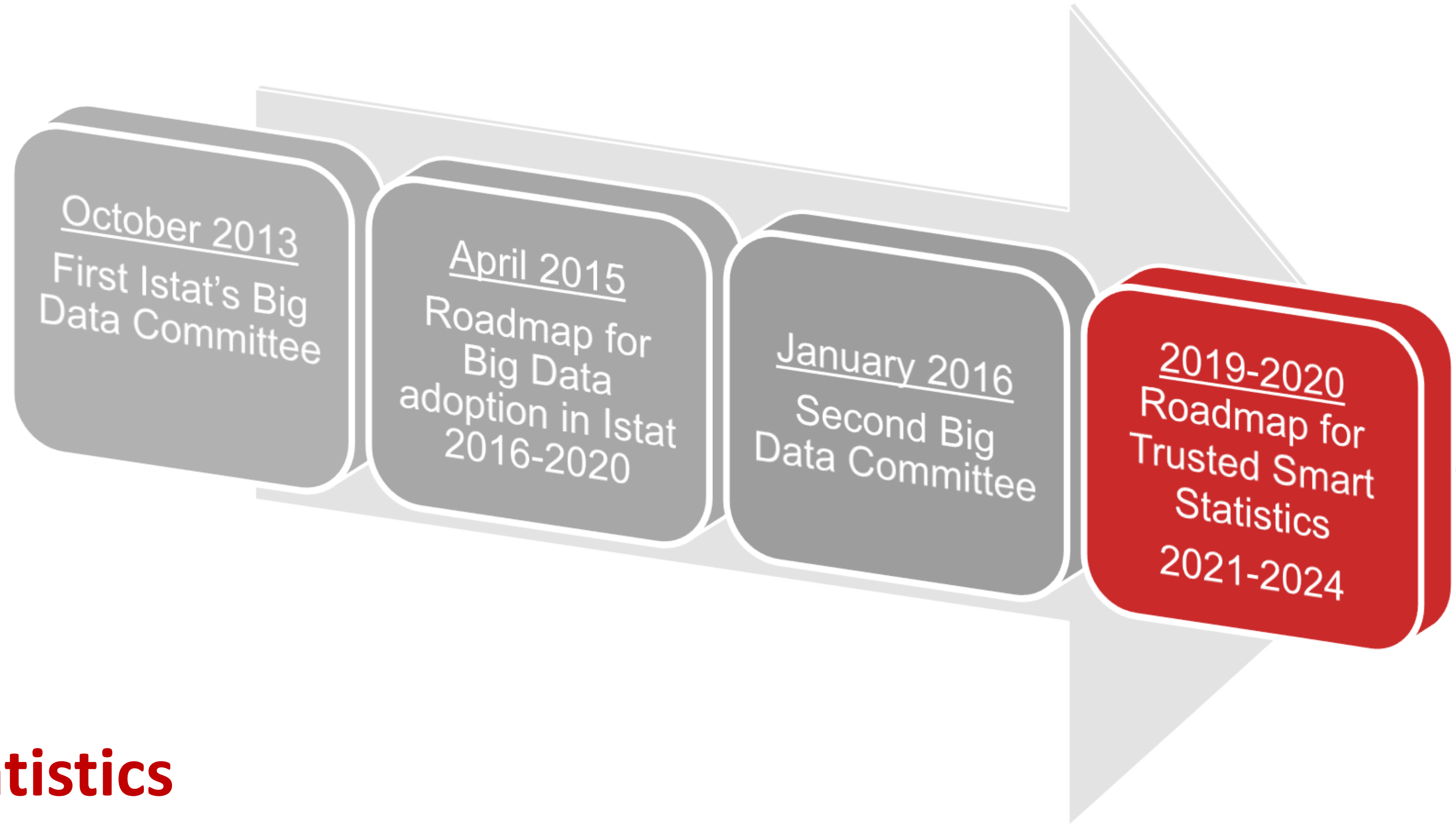


Big Data @Istat: How

○ Istat's strategic context:

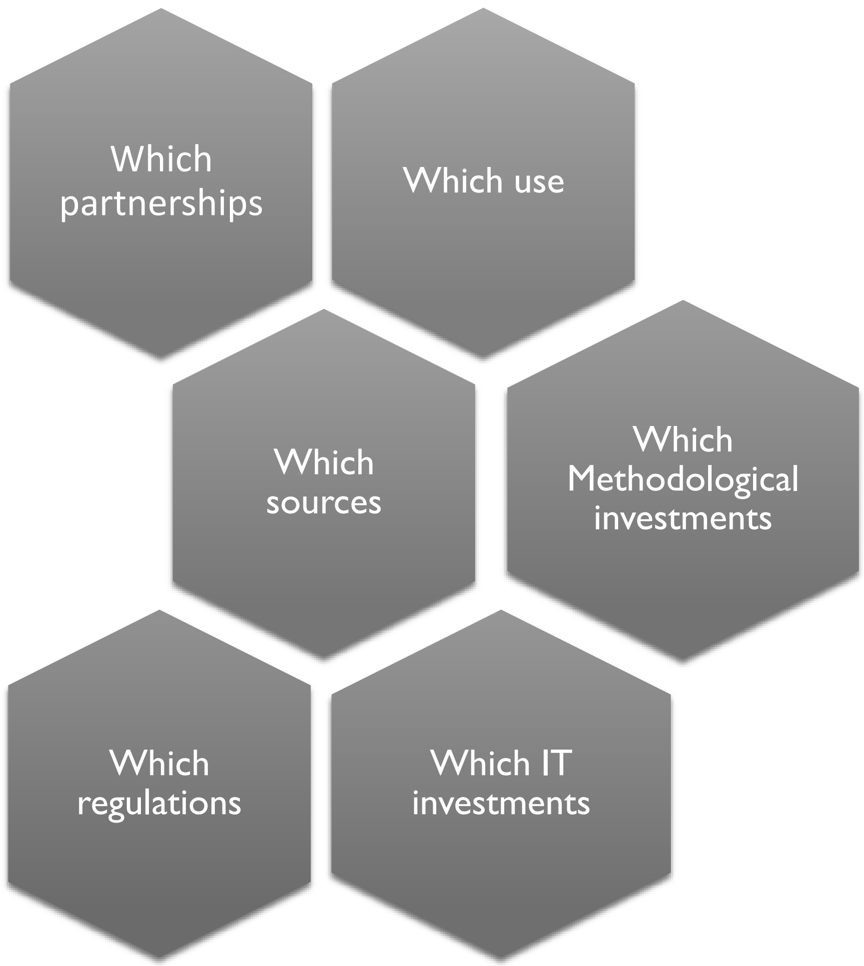
The use of Big Data in Official Statistics requires **methodological**, **technological** and **organizational** investments.

Starting from 2013, Istat created a Big Data Committee responsible of the Big Data strategy…



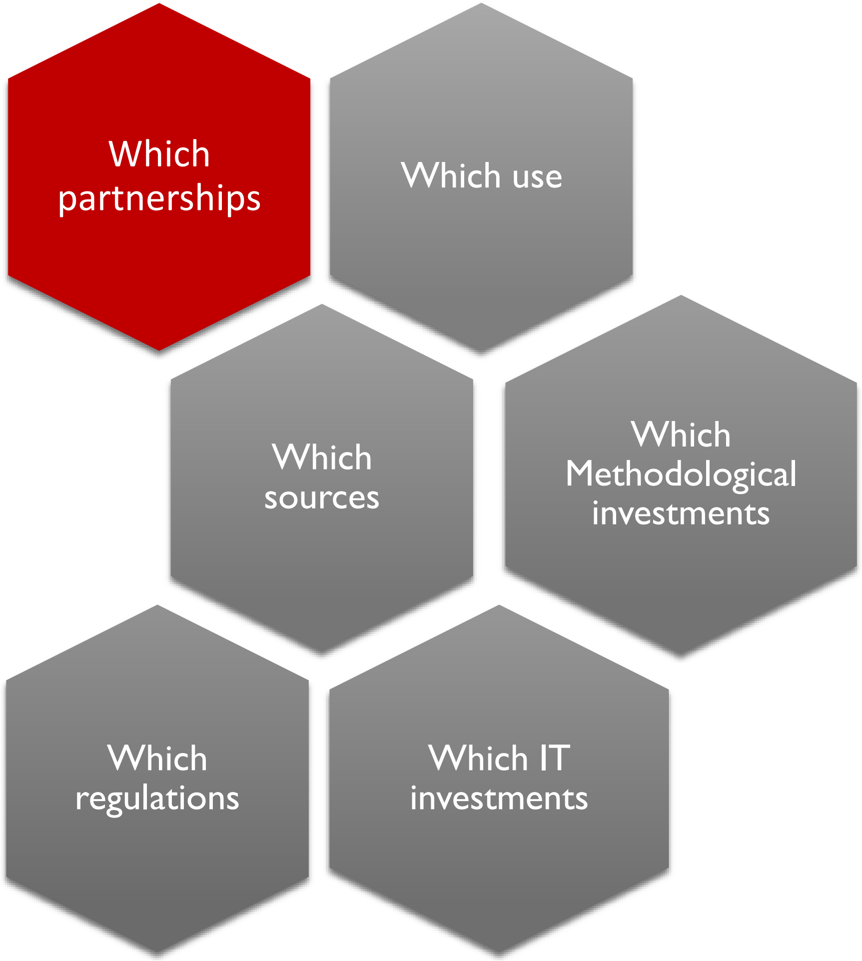
…towards **Trusted Smart Statistics**

Big Data @Istat: How – Operational Dimensions



Big Data @Istat: How – Operational Dimensions

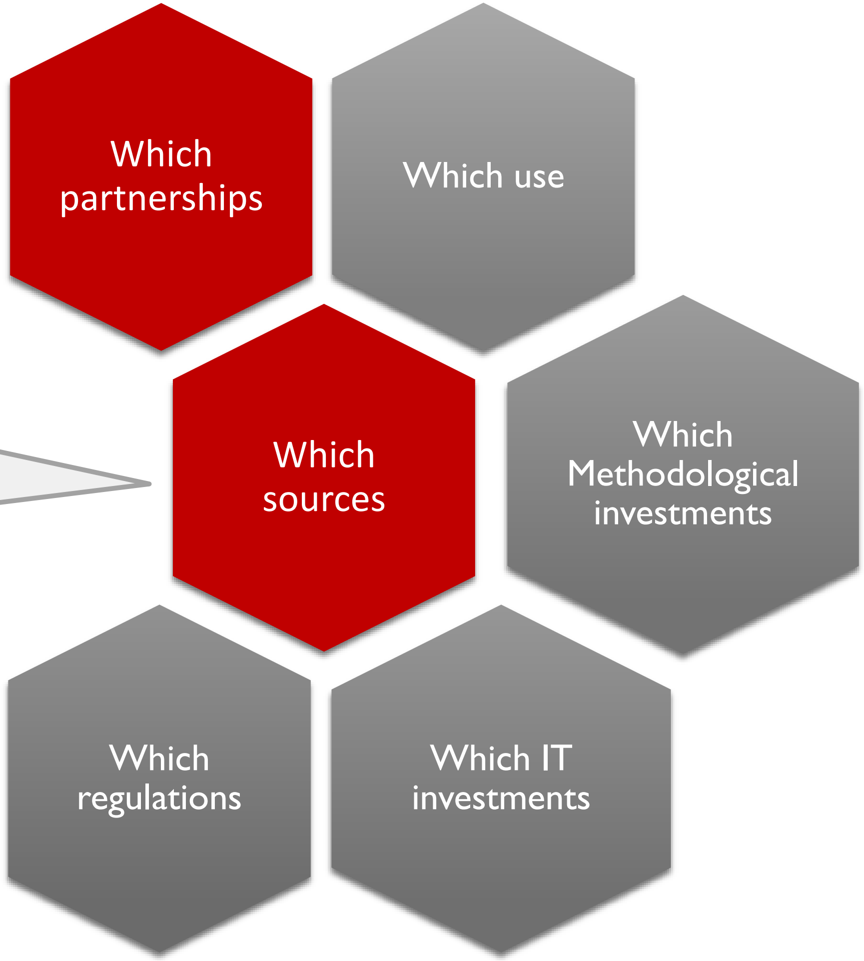
- Mobile Network Operators
- Academic partnerships
- Private Companies (e.g. Nielsen)
- …



Big Data @Istat: How – Operational Dimensions

- Mobile Network Operators
- Academic partnerships
- Private Companies (e.g. Nielsen)
- …

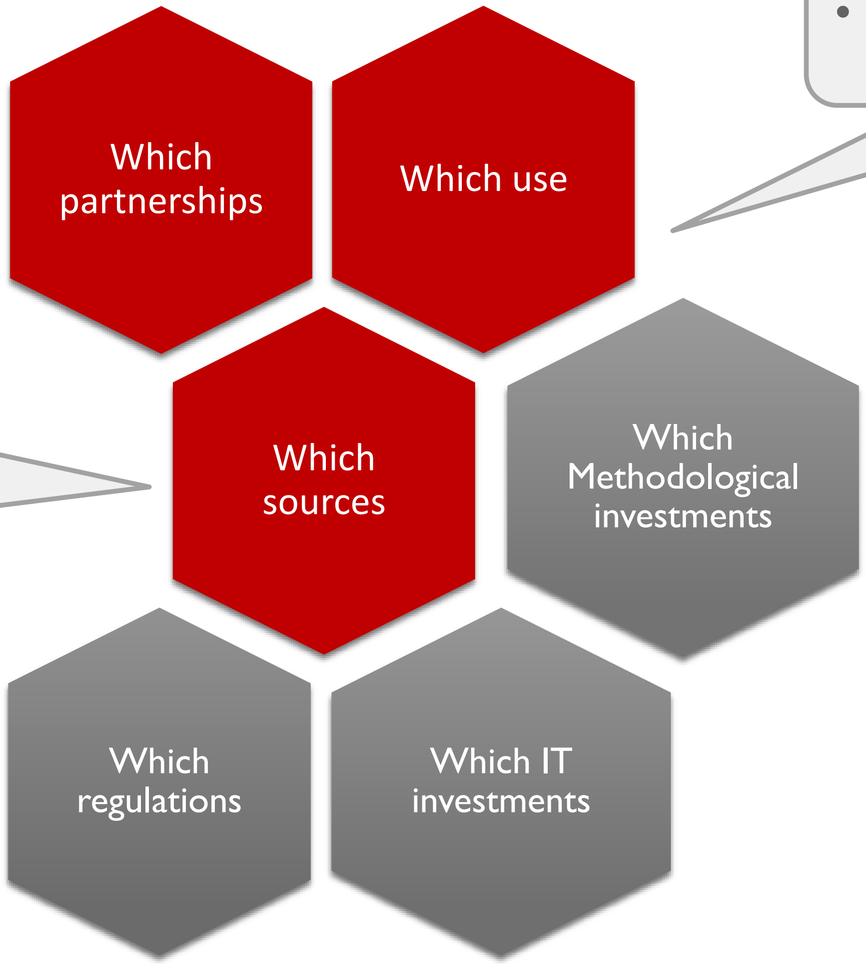
- Web data
- Twitter Data
- Satellite Imagery
- MNO Data
- Mobile Device Sensors
- Scanner Data
- AIS Data
- …



Big Data @Istat: How – Operational Dimensions

- Mobile Network Operators
- Academic partnerships
- Private Companies (e.g. Nielsen)
- …

- Web data
- Twitter Data
- Satellite Imagery
- MNO Data
- Mobile Device Sensors
- Scanner Data
- AIS Data
- …



- **Experimental Statistics**
- Enhanced Official Statistics Products

Big Data @Istat: How – Operational Dimensions

- Mobile Network Operators
- Academic partnerships
- Private Companies (e.g. Nielsen)
- …

Which partnerships

Which use

- **Experimental Statistics**
- Enhanced Official Statistics Products

- Web data
- Twitter Data
- Satellite Imagery
- MNO Data
- Mobile Device Sensors
- Scanner Data
- AIS Data
- …

Which sources

Which Methodological investments

- GDPR
- Data Act
- …

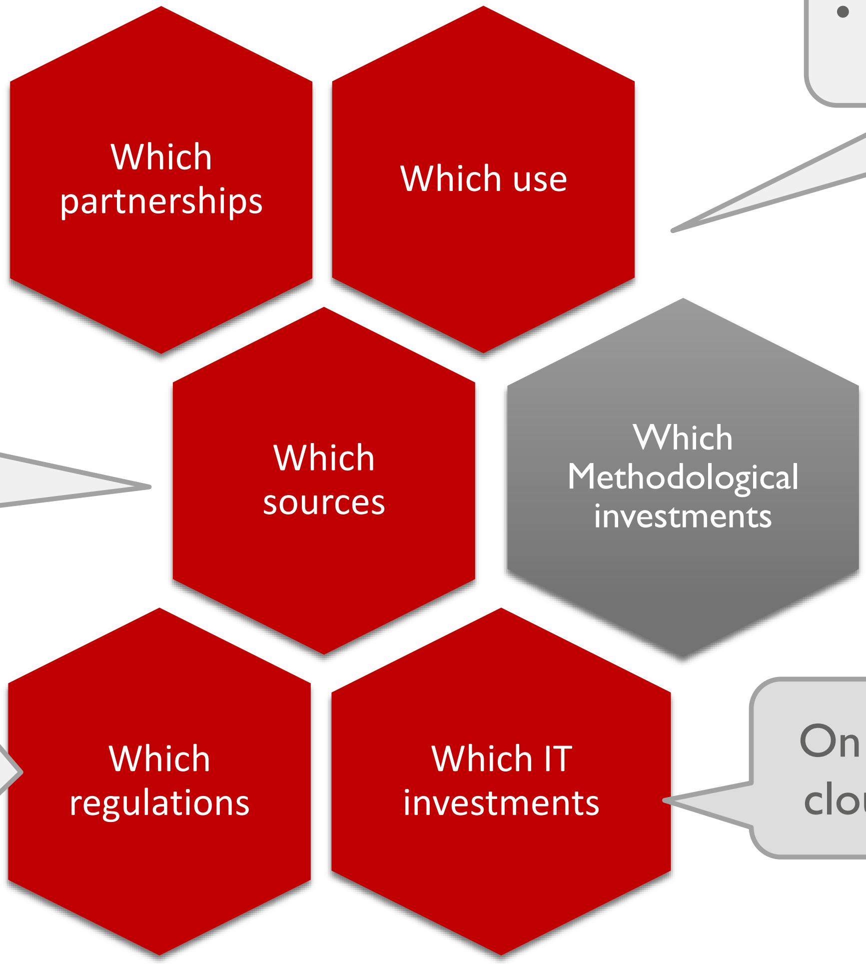
Which regulations

Which IT investments

Big Data @Istat: How – Operational Dimensions

- Mobile Network Operators
- Academic partnerships
- Private Companies (e.g. Nielsen)
- …

- **Experimental Statistics**
- Enhanced Official Statistics Products

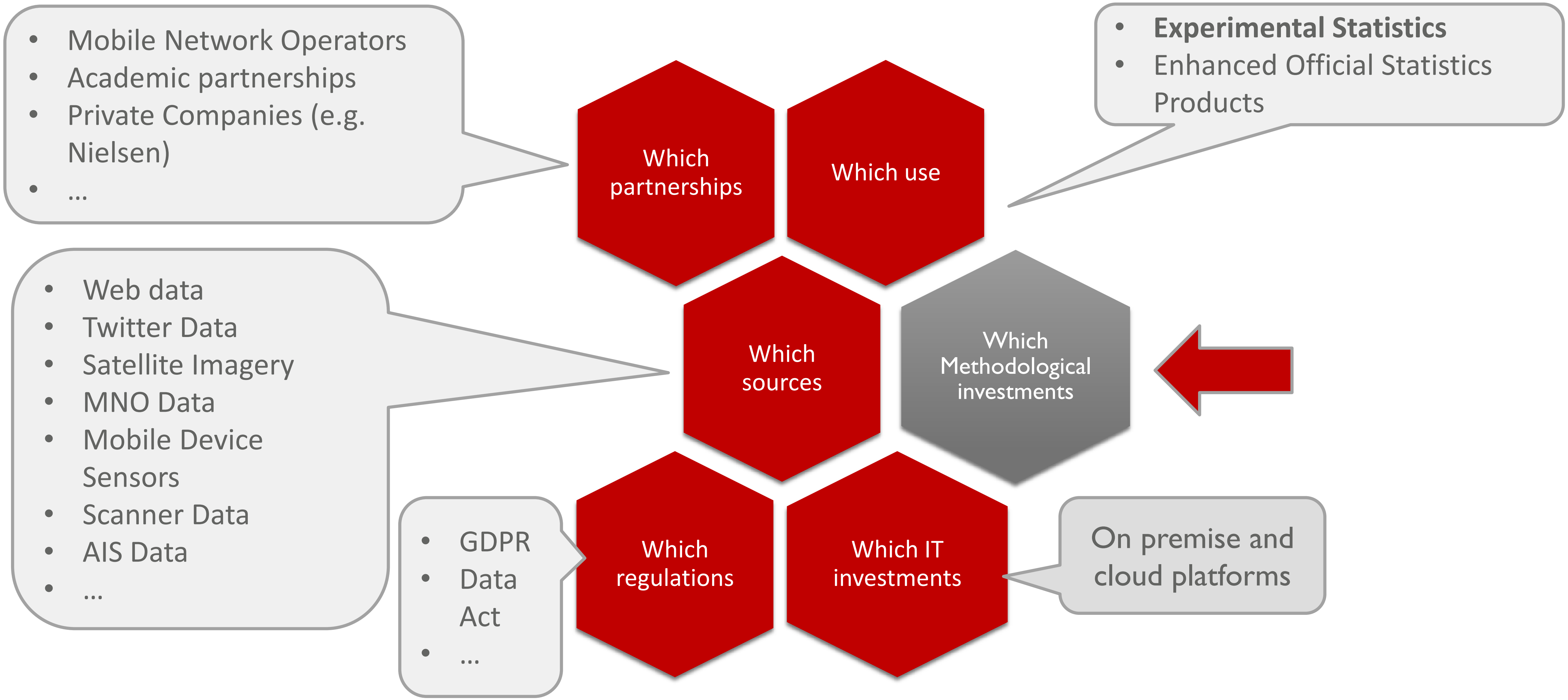


- Web data
- Twitter Data
- Satellite Imagery
- MNO Data
- Mobile Device Sensors
- Scanner Data
- AIS Data
- …

- GDPR
- Data Act
- …

On premise and cloud platforms

Big Data @Istat: How – Operational Dimensions



Istat

Big Data @Istat: What – Methodology

- **Dealing with Big Data heterogeneity**
  - New **data preparation** pipelines
    - Text
    - Images
  - New **inference** paradigm
    - **Machine learning**

Which Methodological investments

- **Dealing with access to external (private) sources**
  - Privacy preserving methods
  - Web scraping

Big Data Projects @Istat

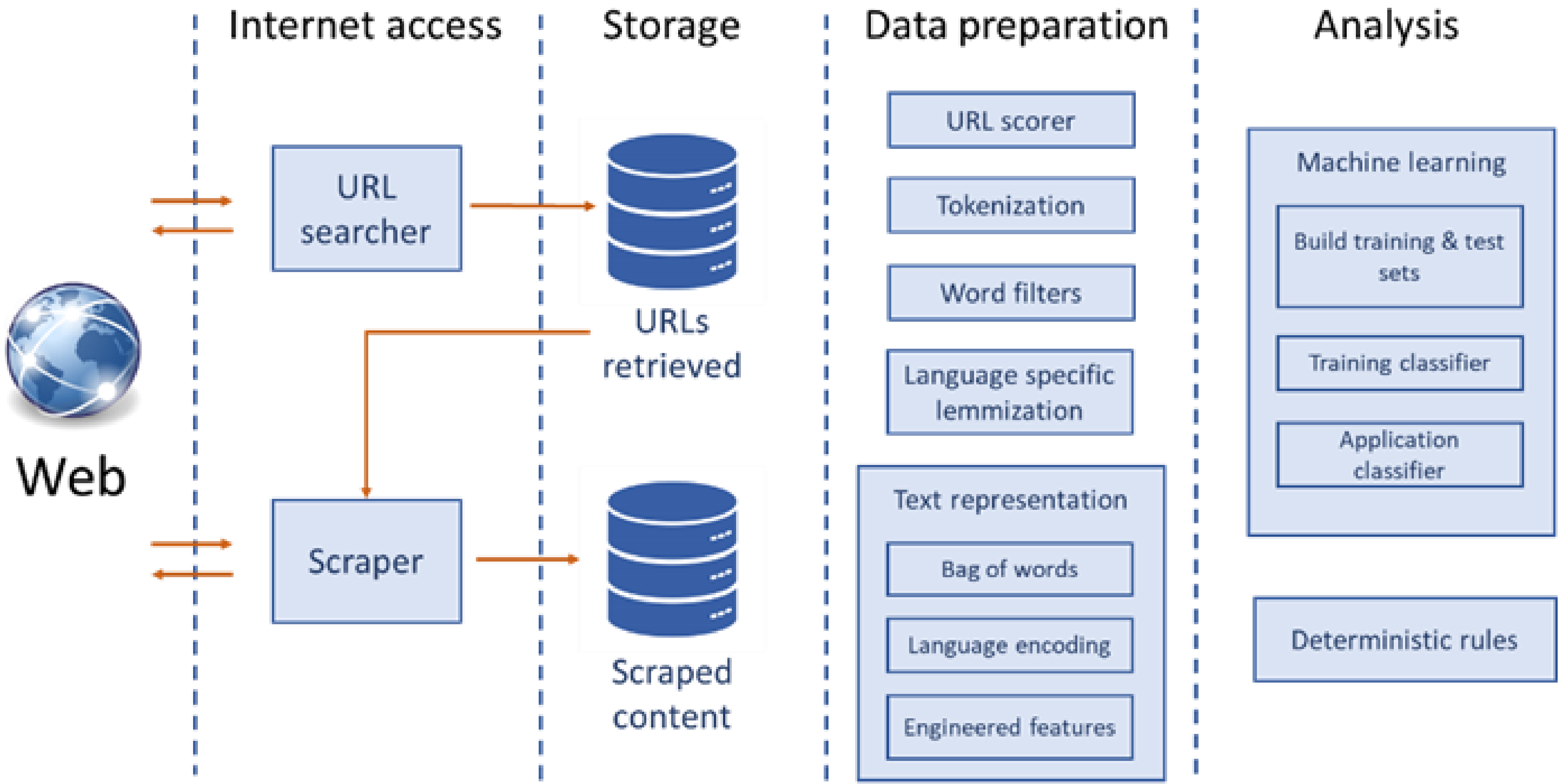
Text processing pipelines: Enterprise Characteristics

GOALS:

- The main goal of the project is the estimation of enterprises characteristics starting from the **web scraping of enterprises websites** (e.g., web ordering, job vacancy advertisement, link to social media)
- We implemented an algorithm that allows to predict enterprises characteristics (supervised machine learning model). The model is trained with survey data serving as training set for the machine learning task
- Simulations have demonstrated that these **new estimates are comparable to survey-based estimates**

Text processing pipelines: Enterprise Characteristics

Generic pipeline for processing textual data from enterprise websites:



Text processing pipelines: Enterprise Characteristics

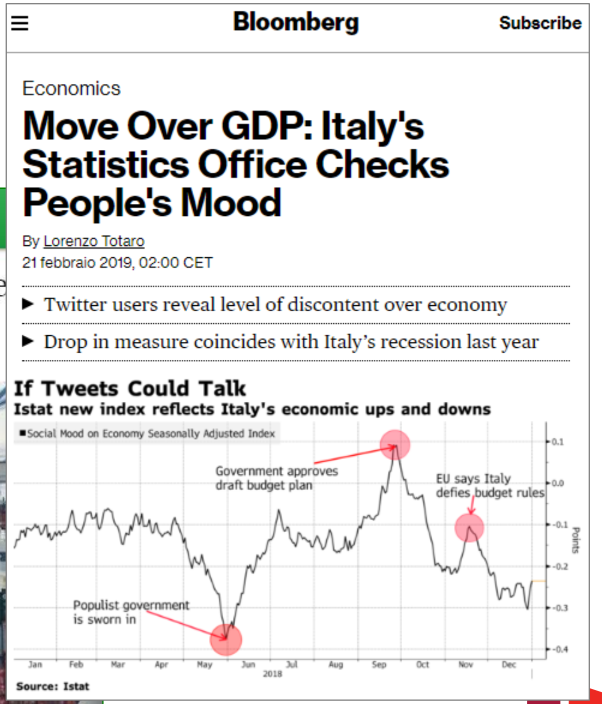
Open challenges:

- The described project faces (some aspects of) the important issue of integrating a Big Data source with survey data. In the project, survey data are used as a training set of a Machine Learning classifier executed on Web extracted data
- Recently, in (Pratesi et al., 2022), more complex data integration methods are used to reduce the bias by combining a probability and a non-probability sample through a vector of common auxiliary variables, as an extension of (Kim & Wang, 2019).

Text processing pipelines: Social Mood

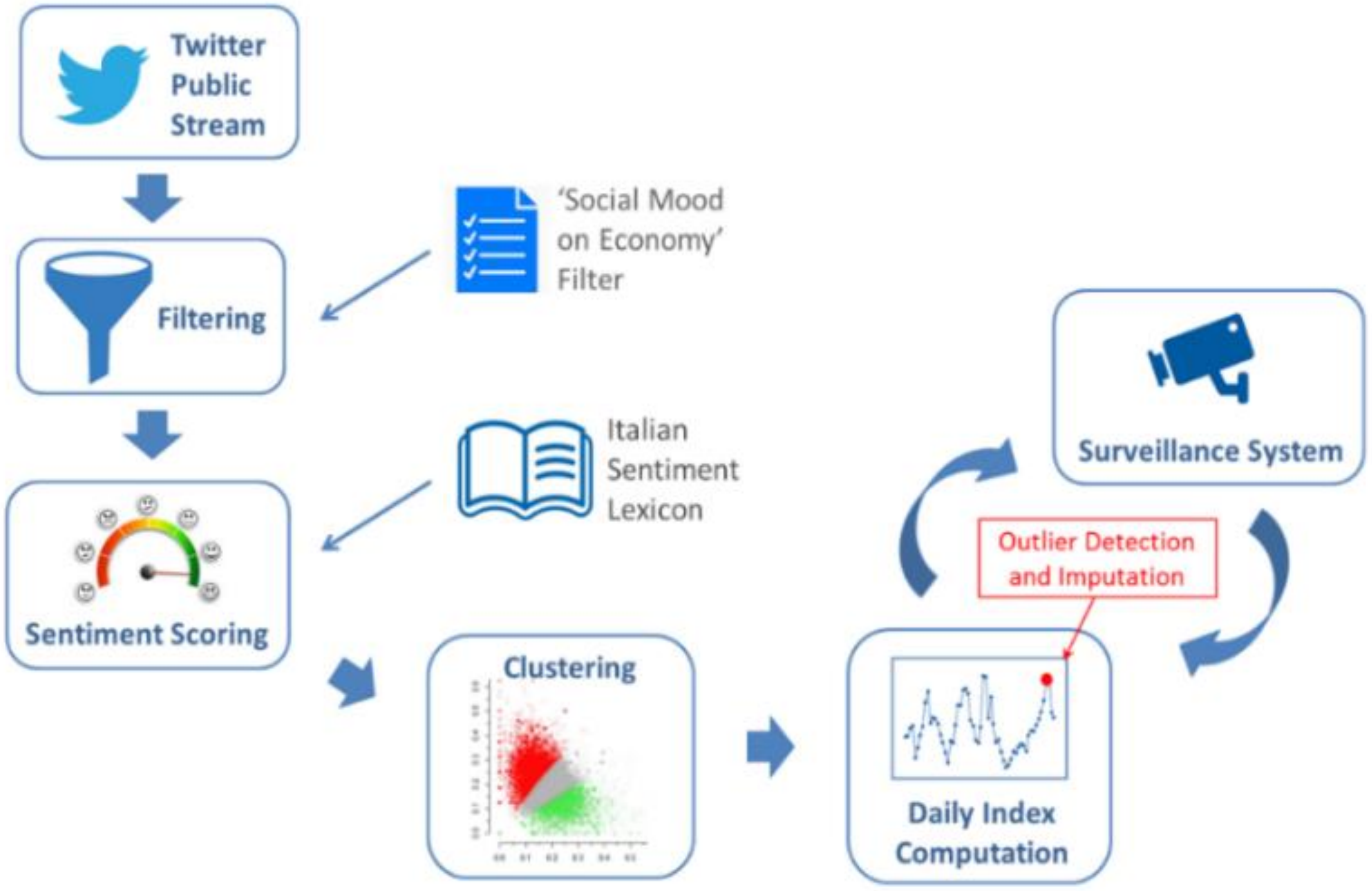
GOALS:

- The Social Mood on Economy Index (SMEI) is an **experimental statistic** published by Istat since 2018. It provides daily measures of the Italian sentiment on the economy, these measures derived from samples of public tweets in Italian language captured in real time
- Data collection started in February 2016 and has been active since then almost without interruptions
  - The dissemination of the new index attracted significant interest from the media (both traditional and online)
  - Receptions were predominantly positive (the overwhelming majority praising Istat's openness to innovation)
  - But few skeptical comments too!



Text processing pipelines: Social Mood

Pipeline to produce the Social Mood on Economy Index



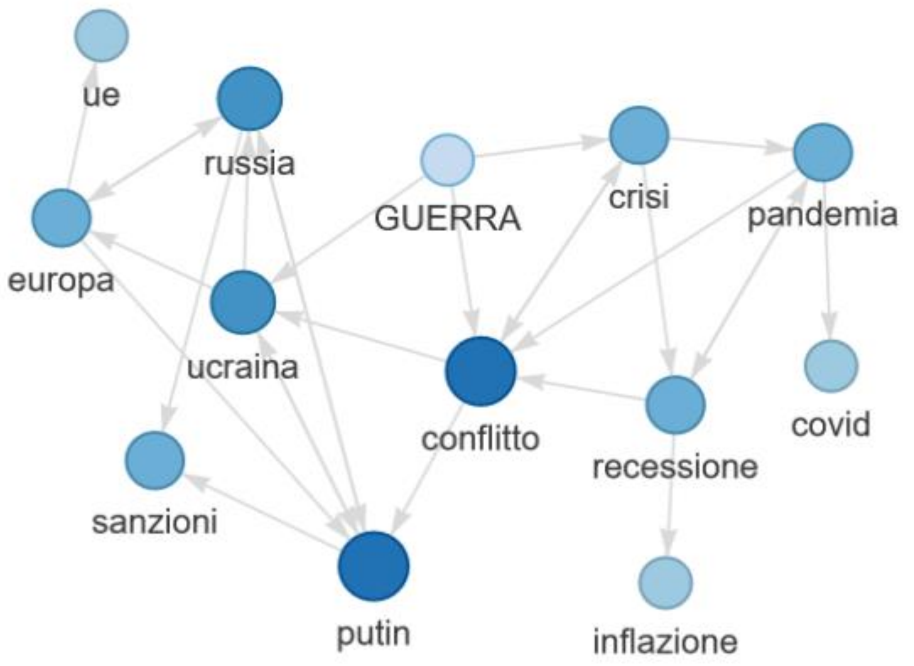
Text processing pipelines: Social Mood

Open challenges:

- In relation to SMEI, there are two major research directions that we would like to explore, namely:
  - (i) evaluation of the quality of Twitter's filters
  - (ii) improvement of the index interpretability

WordEmBox

To evaluate the quality of the filter keywords we have exploited Word Embeddings (WE) methods. To this aim we used WordEmBox, an ad-hoc tool developed by Istat aiming at exploring WE spaces



Graph analysis for word "GUERRA" with the WordEmBox

Image processing pipelines: Land Cover

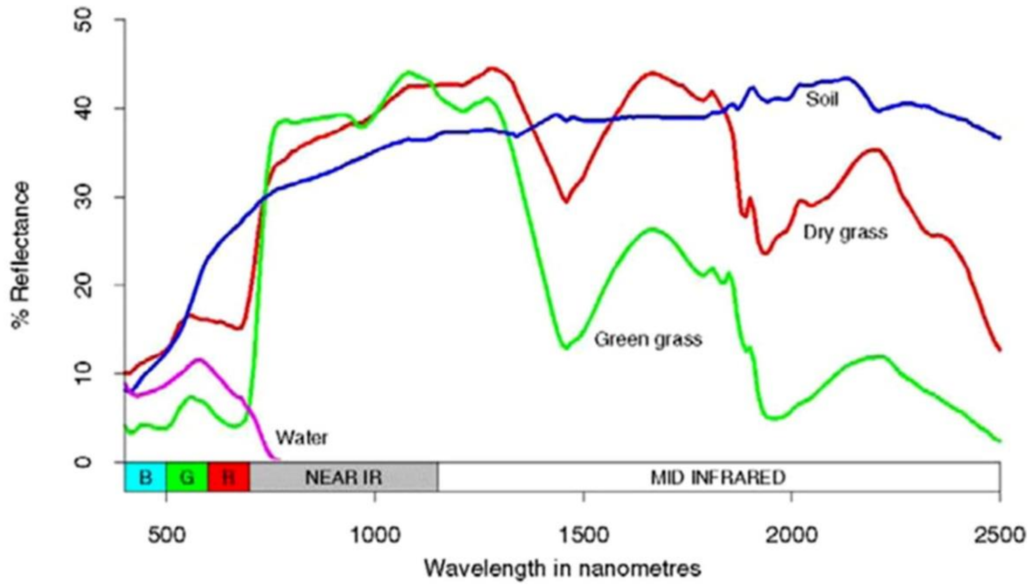
GOALS:

Land Cover (LC) statistics and maps are a very important statistical product. As they require a big effort to be created, the idea is to build a set of algorithms to process satellite images in order to generate:

- Automatic Land Cover **Estimates**
- Automatic Land Cover **Maps**

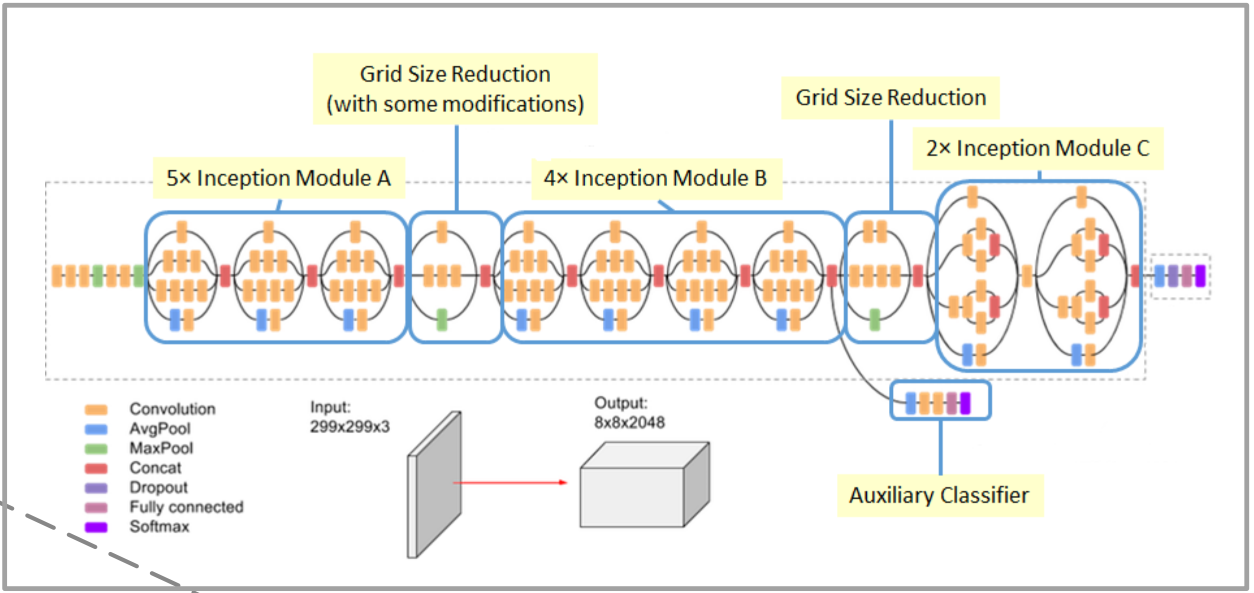
HOW:

- **Standard approach**: Spectral Signature
- **New approach**: Using Deep Learning (**CNN** for classify-and-count and **U-Net** for segmentation)



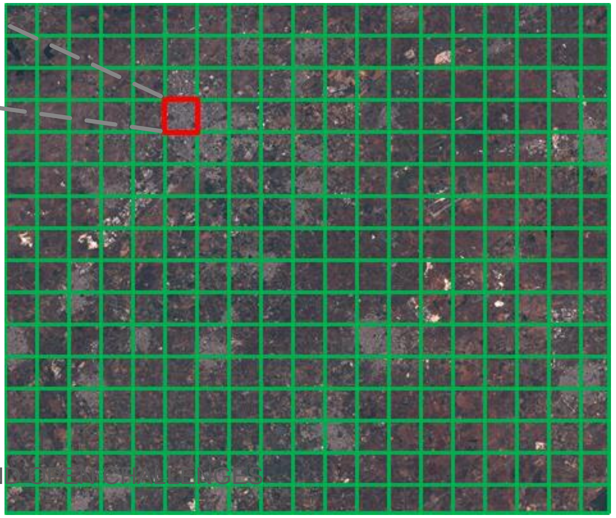
Standard approach: Spectral Signature

Image processing pipelines: Land Cover



RESIDENTIAL

Input Satellite Image



LAND COVER CLASS	AREA SHARE
...	...
RESIDENTIAL	$\frac{45}{16 * 19} \cong 15\%$
...	...

Istat

Image processing pipelines: Land Cover

Results:

The integrated architecture (CNN + U-Net) works very well for all LC classes:

- The U-Net takes care of LC classes “**River**” and “**Highway**”
- The CNN copes with all the other LC classes
- Partial LC maps produced by 1) and 2) are merged to yield a final complete LC map

Open challenges:

- One of the major problems in automated Land Cover (LC) estimation project is the **lack of a benchmark to validate the algorithm**, according to the chosen resolution and type of classification
- Suitable training dataset, compatible with the requested resolution for output. **Integration of input data with administrative sources** (e.g., data from regional technical charts, cadastral maps, and agricultural census)

Improving data dissemination pipeline: TERRA

Istat's experience at European Big Data Hackathon, where we used different data sources traditional (**Comext Data**) and produced by sensors (**Google Mobility**) to analyze the impact of mobility restrictions on import / export

TERRA - imporT ExpoRt networRk Analysis

Istat

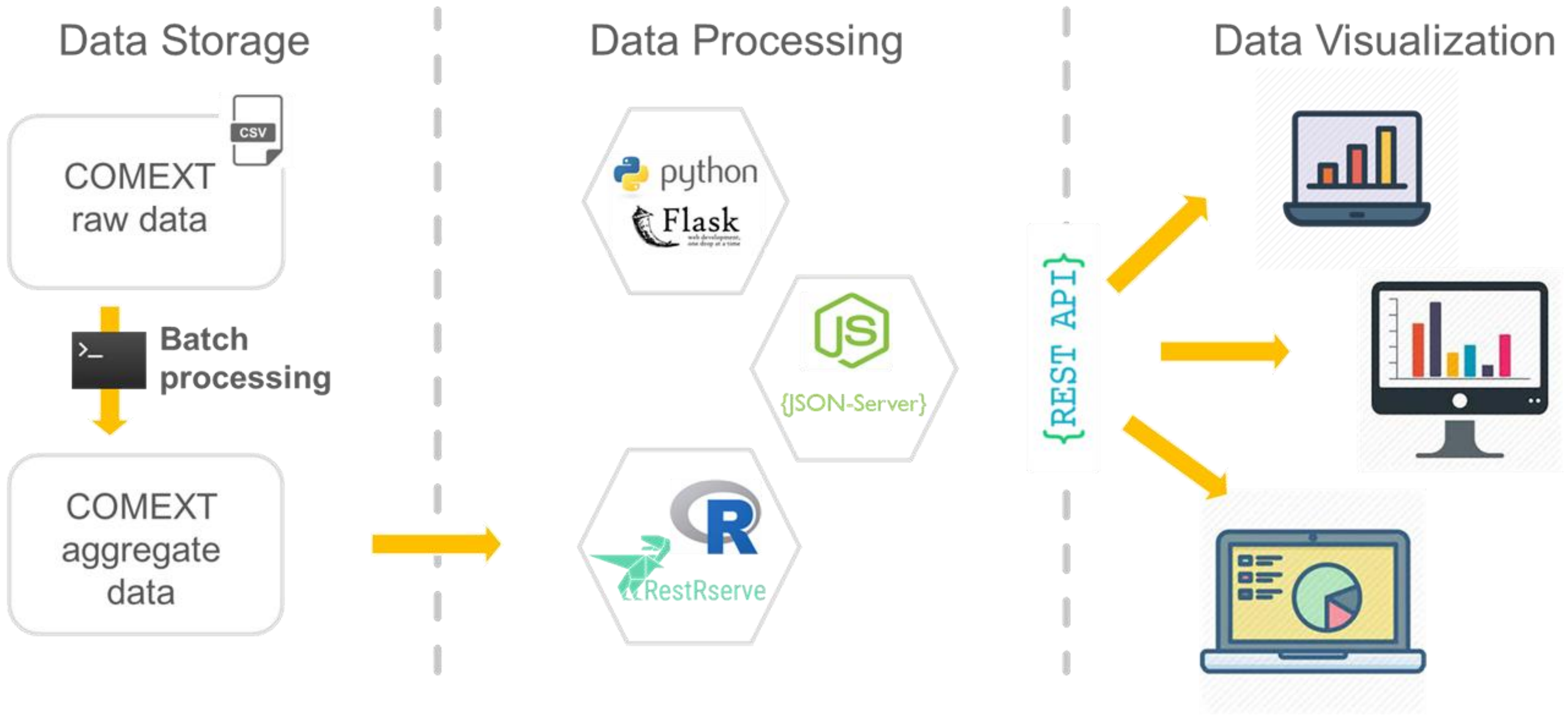
Improving data dissemination pipeline: TERRA

Big Data Hackathon 2021:

- Create an interactive dashboard to predict and simulate international trade relations in a high-resolution network by product and time
  - **Scenario analysis** and support for international trade policies
  - Ability to represent international global exchange networks
  - **Visualization of relationships** for partners, products and means of transportation
  - Analysis at product level as disaggregated as possible
  - **Scenario simulation for transport interruptions**
  - Further suggested analysis: study of the **impact on COVID** products, etc.

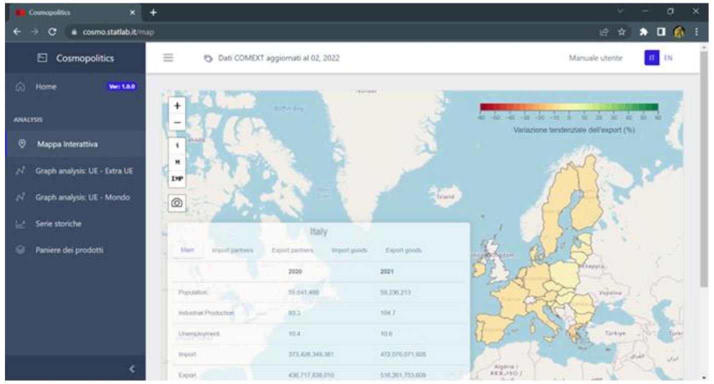
Improving data dissemination pipeline: TERRA

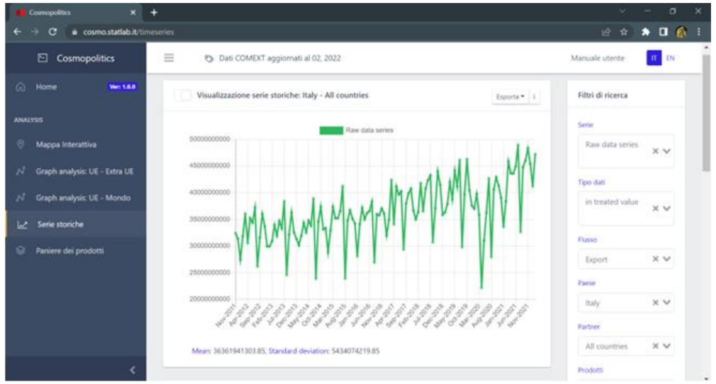
TERRA in action:

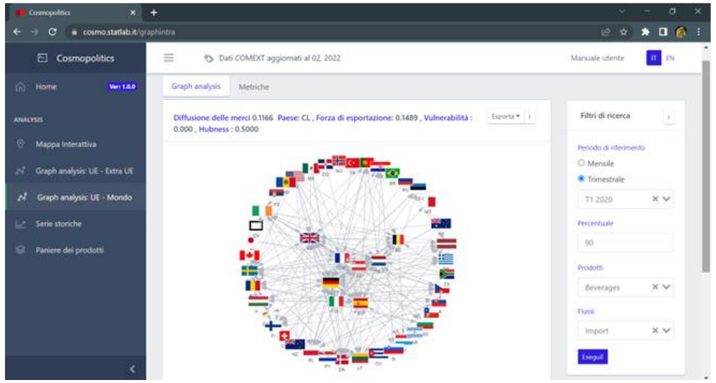


Improving data dissemination pipeline: TERRA

TERRA is online (almost ready for production):







https://www.terra.statlab.it/

Improving data dissemination pipeline: TERRA

Open challenges:

- In its first version, the time series analysis section provided a forecasting up to 6 months a-head of the future international trade flows. This functionality used **Google COVID-19 Mobility Reports** to build a synthetic indicator capable of explaining the level of restriction imposed in each country using the principal component analysis methodology

- **Integration of new data sources (e.g., Stringency Index)**:
- TERRA could provide new tools for performing scenario analysis. Indeed, the time series analysis section could be enriched by the inclusion of a subsection dedicated to one or more open indicators such as the Oxford University Stringency Index

Concluding remarks...

Methodological issues (recap)

Open challenges:

- [**Web scraping**] One of the most challenging aspects of the project concerns the issue of integrating a Big Data source with survey data. In the project, survey data are used as a training set of a Machine Learning classifier executed on Web extracted data
- [**Social Mood**] In relation to SMEI, there are two major research directions that we would like to explore, namely: 1) evaluation of the quality of Twitter's filters; 2) improvement of the index interpretability
- [**Land Cover**] One of the major problems in automated Land Cover (LC) estimation project is the lack of a benchmark to validate the algorithm, according to the chosen resolution and type of classification

Conclusions

- In the last 10 years huge investments on internal capacity building
- Big Data and ML projects are **now a strategic asset in Istat**
  - **Experimental statistics are already there**
  - Production processes based on ML inference are planned in our 2021-2024 **Trusted Smart Statistics roadmap**

Thanks!

MAURO BRUNO| mbruno@istat.it

Istat Istituto Nazionale di Statistica