Tapping into web data for official statistics

Experiences from the ESSnet Web Intelligence Network

European Statistics Day

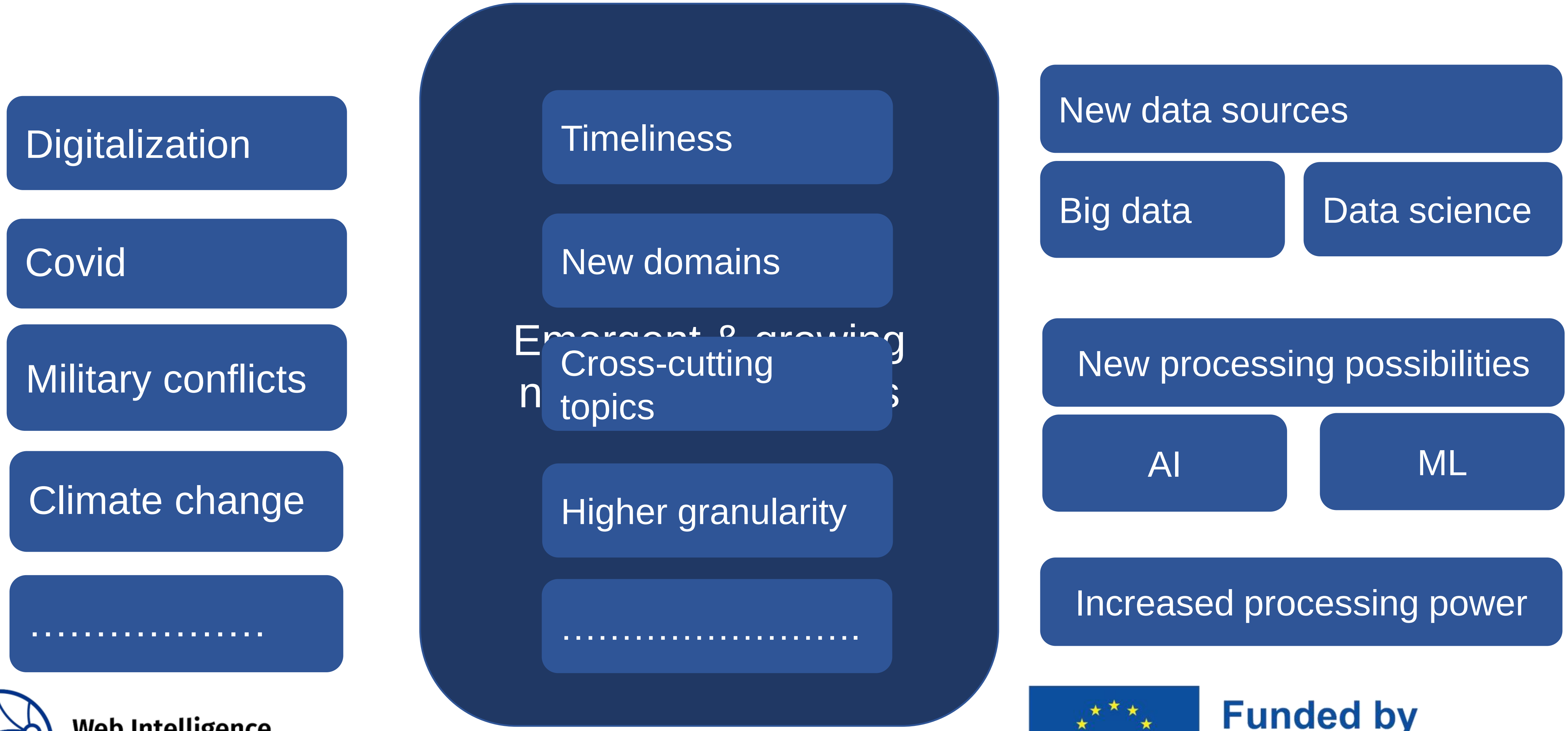
Madrid, October 23, 2023

Dominika Nowak, Statistics Poland

Web Intelligence Network

Funded by the European Union

Constantly changing data landscape



Digitalization

Covid

Military conflicts

Climate change

…………………

Emergent & growing needs

Timeliness

New domains

Cross-cutting topics

Higher granularity

……………………….

New data sources

Big data

Data science

New processing possibilities

AI

ML

Increased processing power

Web Intelligence Network

Funded by the European Union

What it means for official statistics

New methods

New legal arrangements

New work arrangements

New techniques

New cooperation & communication models

Space for experimentation with uncertain results

Changes in infrastructure

New skills

Re-shaping statistical production process

Funded by the European Union

Web Intelligence
Network

Funded by
the European Union

WARNING!

WIN

TSS

WISER

WIP

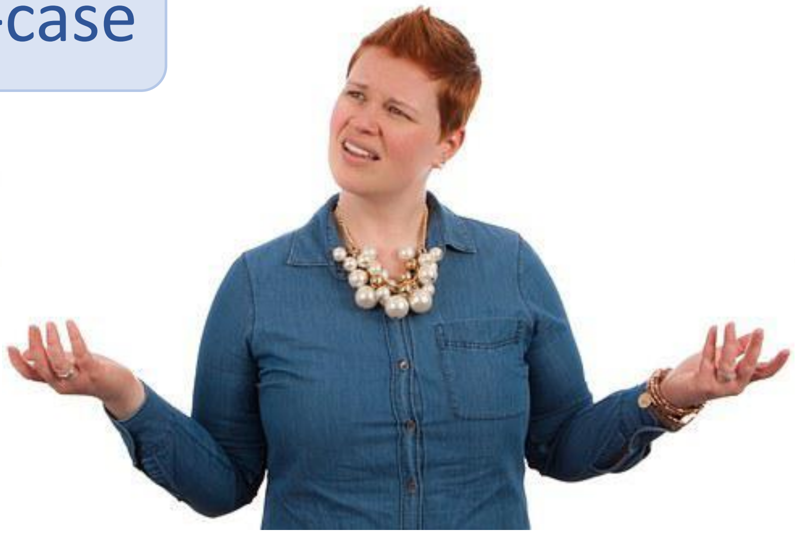
OBEC

Use-case

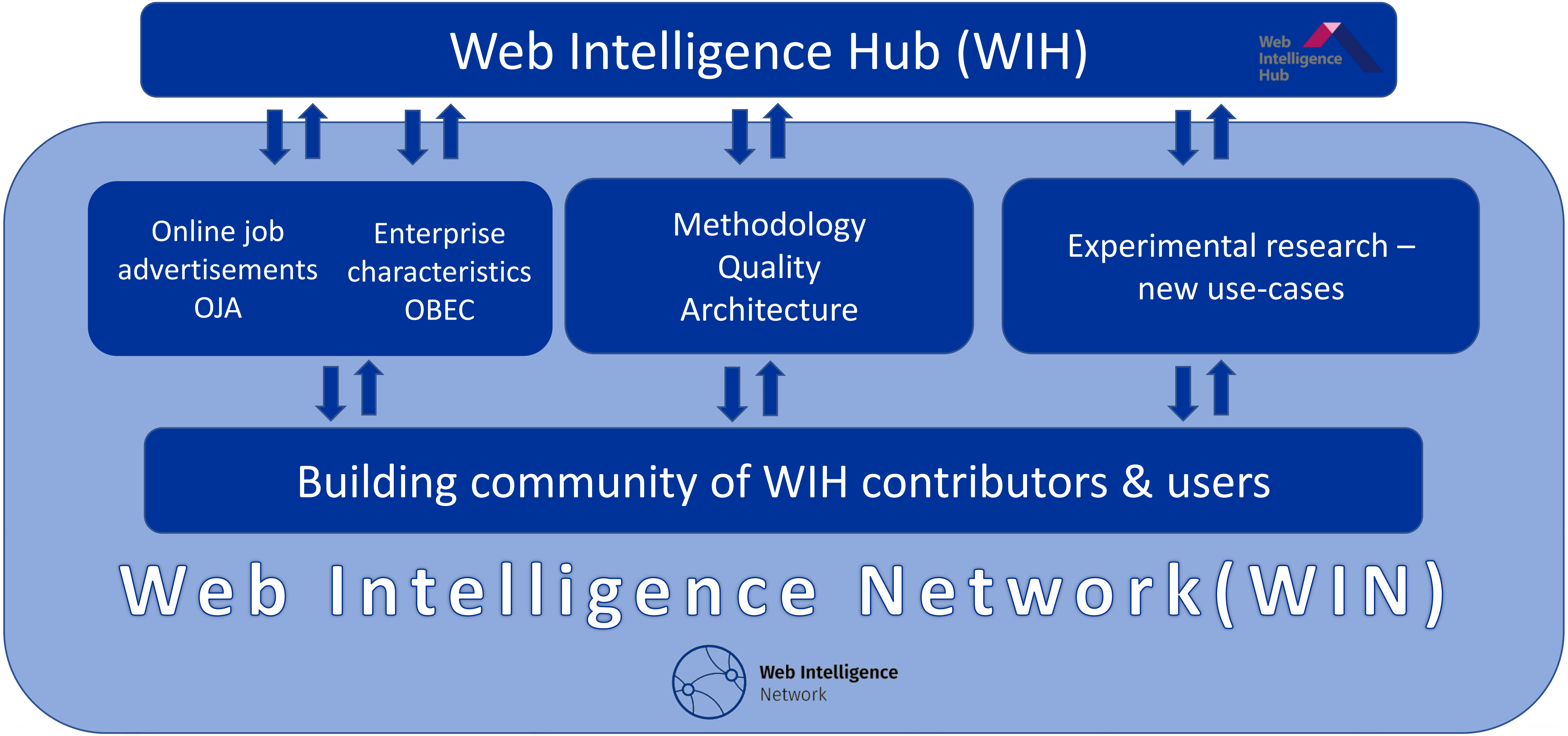
WIH

DataLab

OJA



Funded by the European Union



Web Intelligence Hub (WIH)

Online job advertisements OJA

Enterprise characteristics OBEC

Methodology
Quality
Architecture

Experimental research – new use-cases

Building community of WIH contributors & users

Web Intelligence Network(WIN)

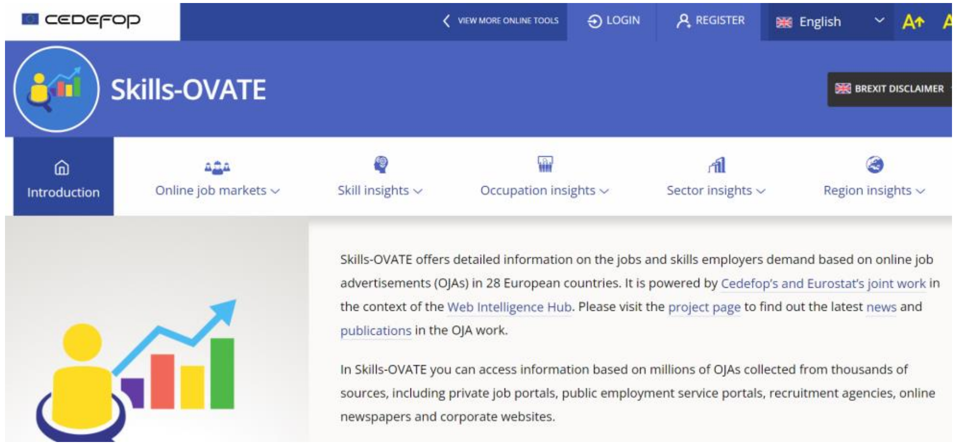
Funded by
the European Union

How it all started…

CEDEFOP

European Centre for the Development
of Vocational Training

eurostat



Skills-OVATE offers detailed information on the jobs and skills employers demand based on online job advertisements (OJAs) in 28 European countries. It is powered by Cedefop's and Eurostat's joint work in the context of the Web Intelligence Hub. Please visit the project page to find out the latest news and publications in the OJA work.

In Skills-OVATE you can access information based on millions of OJAs collected from thousands of sources, including private job portals, public employment service portals, recruitment agencies, online newspapers and corporate websites.

ESSnet

ESSnet Big Data I

ESSnet Big Data II

Webscraping job vacancies

Webscraping enterprise characteristics

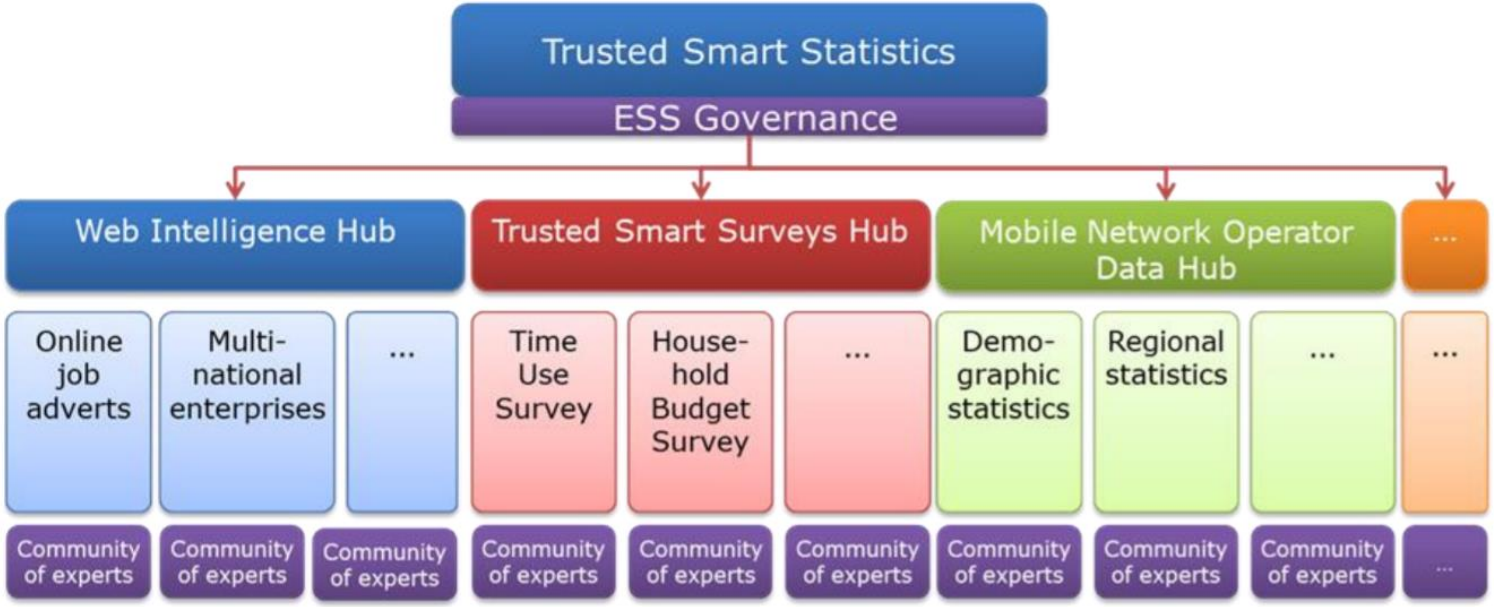
Online job vacancies

Enterprise characteristics

Web Intelligence
Network

What is Web Intelligence Hub (WIH)?

- Started as a concept – evolved towards tangible tools
- Web data acquisition, processing & analysis environment
- Centralized, shared system, pan-European platform



Trusted Smart Statistics

ESS Governance

Web Intelligence Hub: Online job adverts; Multi-national enterprises; …

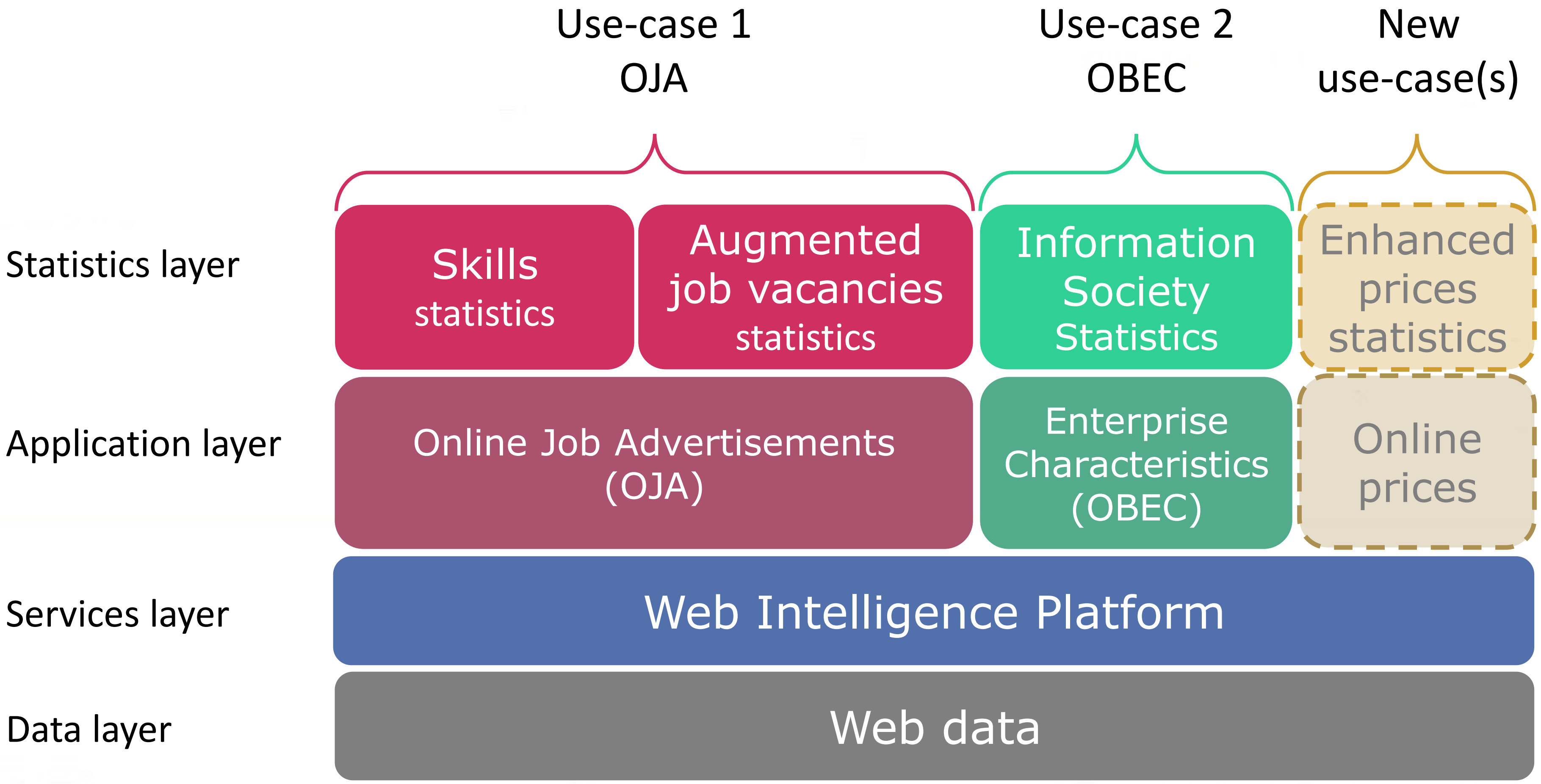
Trusted Smart Surveys Hub: Time Use Survey; House-hold Budget Survey; …

Mobile Network Operator Data Hub: Demo-graphic statistics; Regional statistics; …

…: …

Community of experts | Community of experts | Community of experts | Community of experts | Community of experts | Community of experts | Community of experts | Community of experts | Community of experts | …

What is Web Intelligence Hub (WIH)?



Use-case 1 OJA

Use-case 2 OBEC

New use-case(s)

Statistics layer: Skills statistics · Augmented job vacancies statistics · Information Society Statistics · Enhanced prices statistics

Application layer: Online Job Advertisements (OJA) · Enterprise Characteristics (OBEC) · Online prices

Services layer: Web Intelligence Platform

Data layer: Web data

Web Intelligence Network

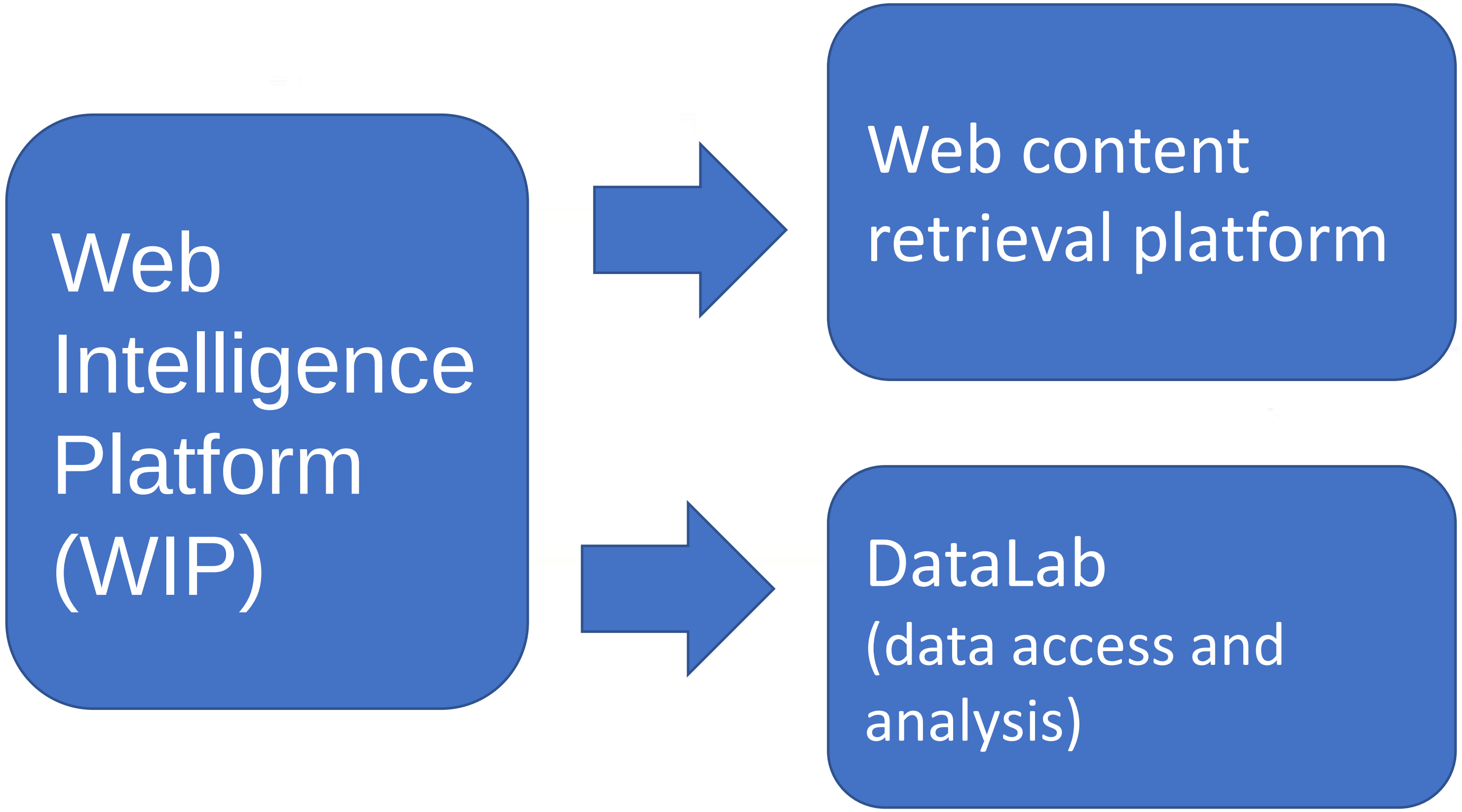
Funded by the European Union

Why a shared system for web-based statistics?

- Different capacity of NSIs across Europe to use web data
- Different competency levels, scarcity of data science skills
- Infrastructure with big data capabilities required
- More efficient use of resources

Funded by the European Union

What WIH concept has materialized into to date



Web Intelligence Platform (WIP)

Web content retrieval platform

DataLab
(data access and analysis)

Funded by the European Union

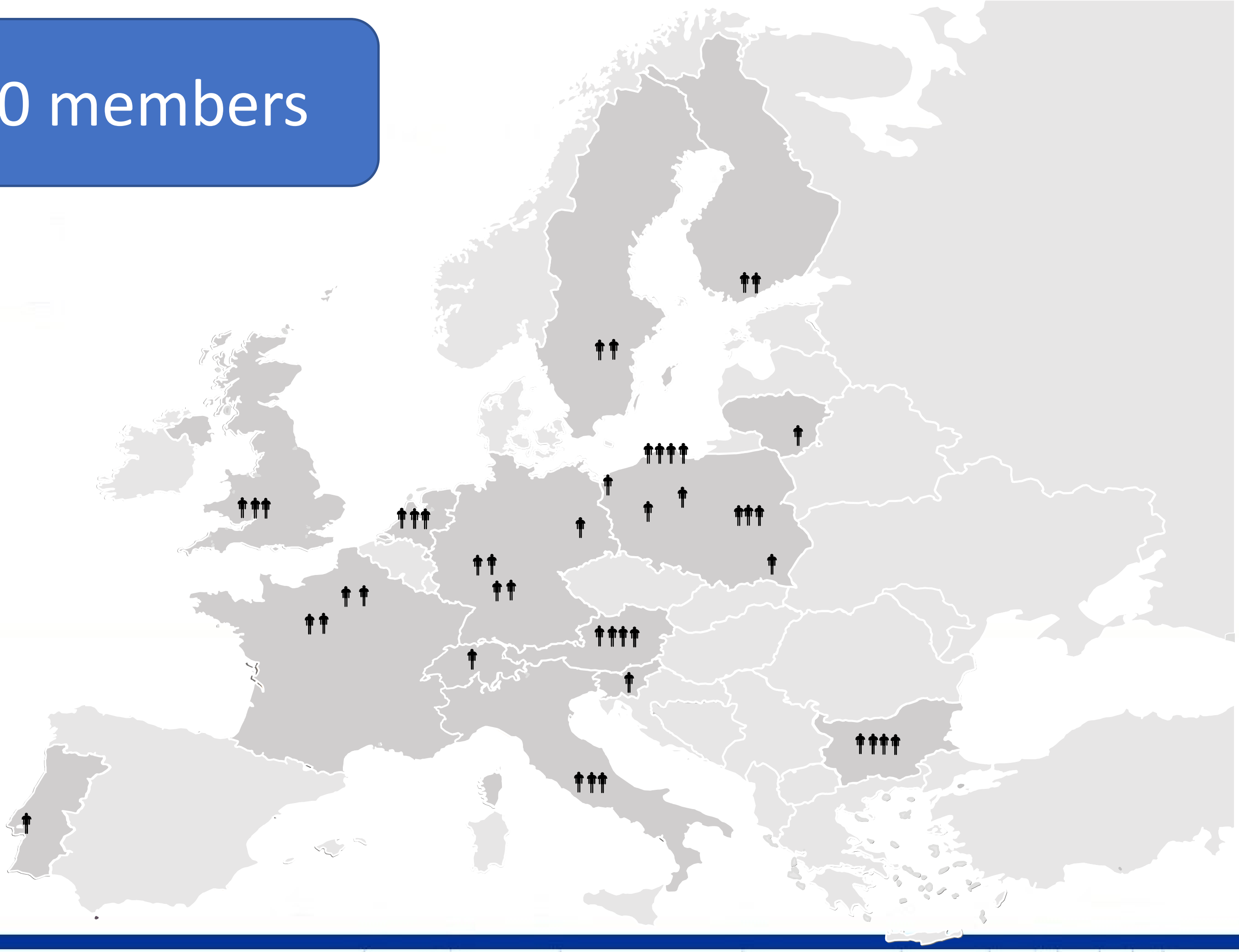
Web Intelligence Network (WIN)

14 countries, 17 organizations, 100 members

Contribute to the development of the WIH

Reach out to **all ESS countries**

Use web data, use the WIH



Web Intelligence Network

Funded by the European Union

How these two work together

Web Intelligence
Network

Web
Intelligence
Hub

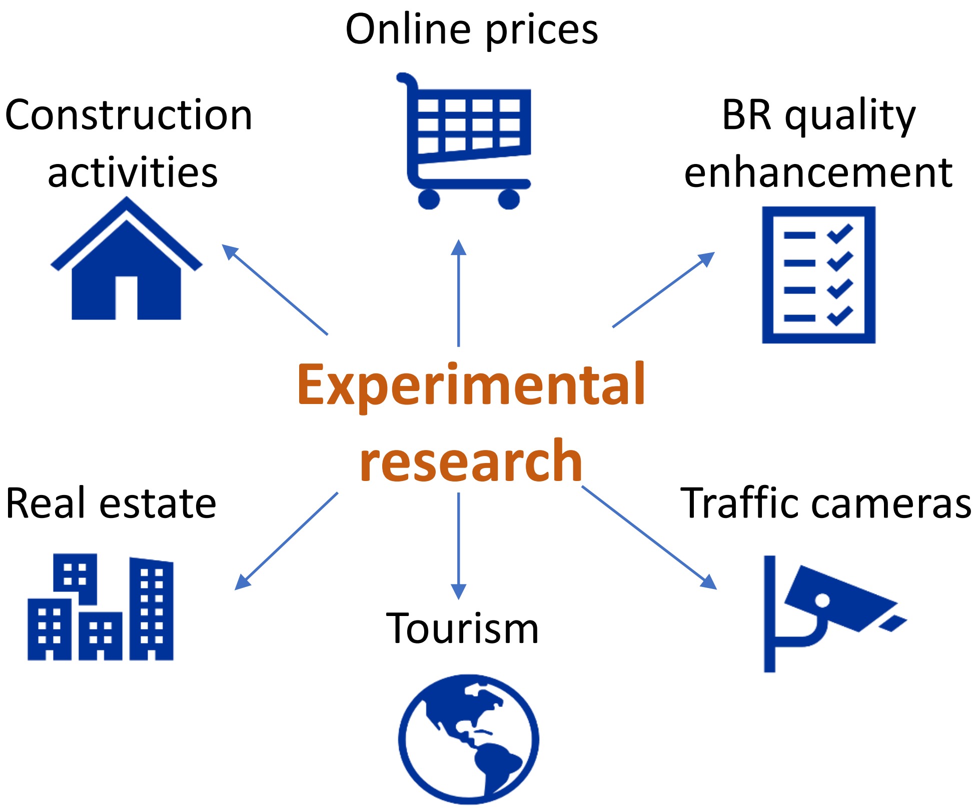
Funded by
the European Union

What topics WIN is looking to

Most mature use-cases

Online job advertisements
OJA

Enterprise characteristics
OBEC



Experimental research

Online prices

Construction activities

BR quality enhancement

Real estate

Tourism

Traffic cameras

Web Intelligence
Network

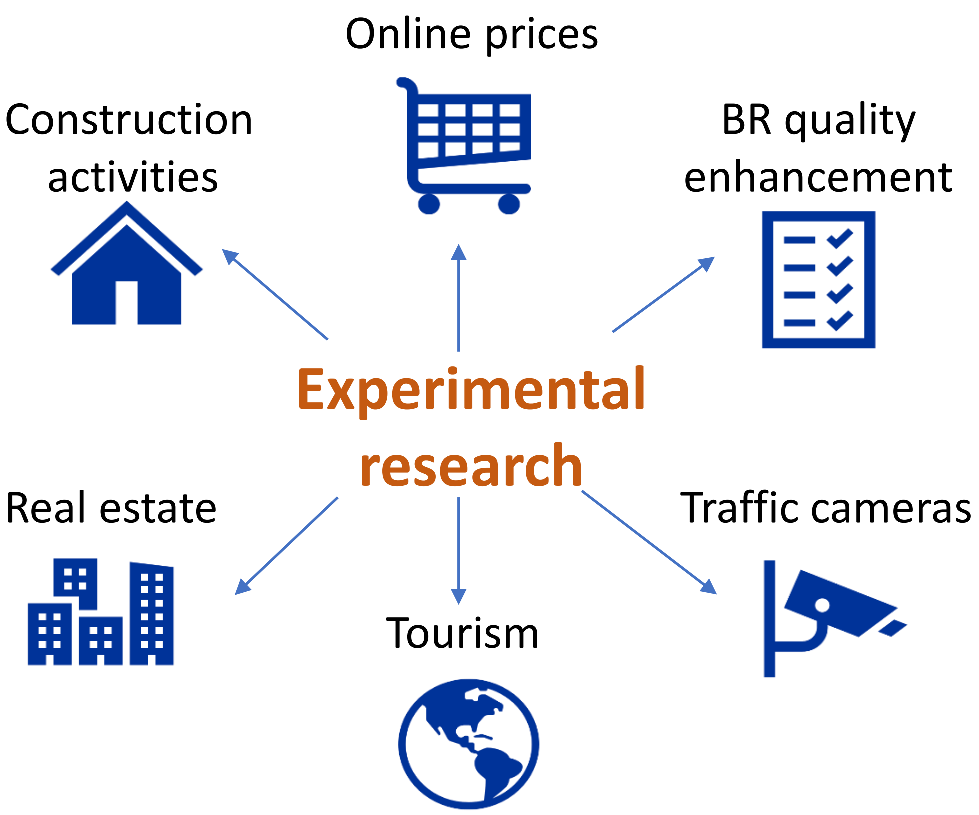
Funded by
the European Union

What topics WIN is looking to

Most mature use-cases

Online job advertisements
OJA

Enterprise characteristics
OBEC



Web Intelligence
Network

Funded by
the European Union

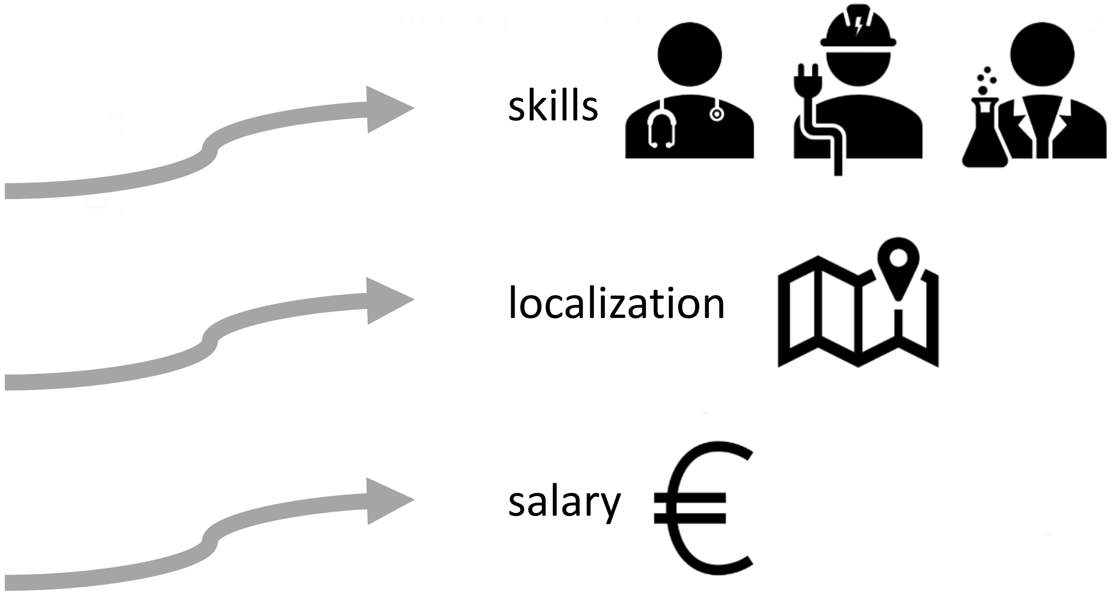
Online Job Advertisements (OJA)

Requirment:

- Lorem ipsum
- Dolor sit amet
- Cras gravida
- Vestibulum ante
- Lorem ipsum
- Dolor sit amet
- Cras gravida
- Vestibulum ante

Job Description:

- Lorem ipsum
- Dolor sit amet
- Cras gravida
- Vestibulum ante
- Lorem ipsum
- Dolor sit amet
- Cras gravida
- Vestibulum ante



Web Intelligence Network

Funded by the European Union

Online job advertisements (OJA)

- Data acquisition and analysis system already in place – data centrally scraped by Eurostat/ Cedefop, accessible by NSIs via DataLab

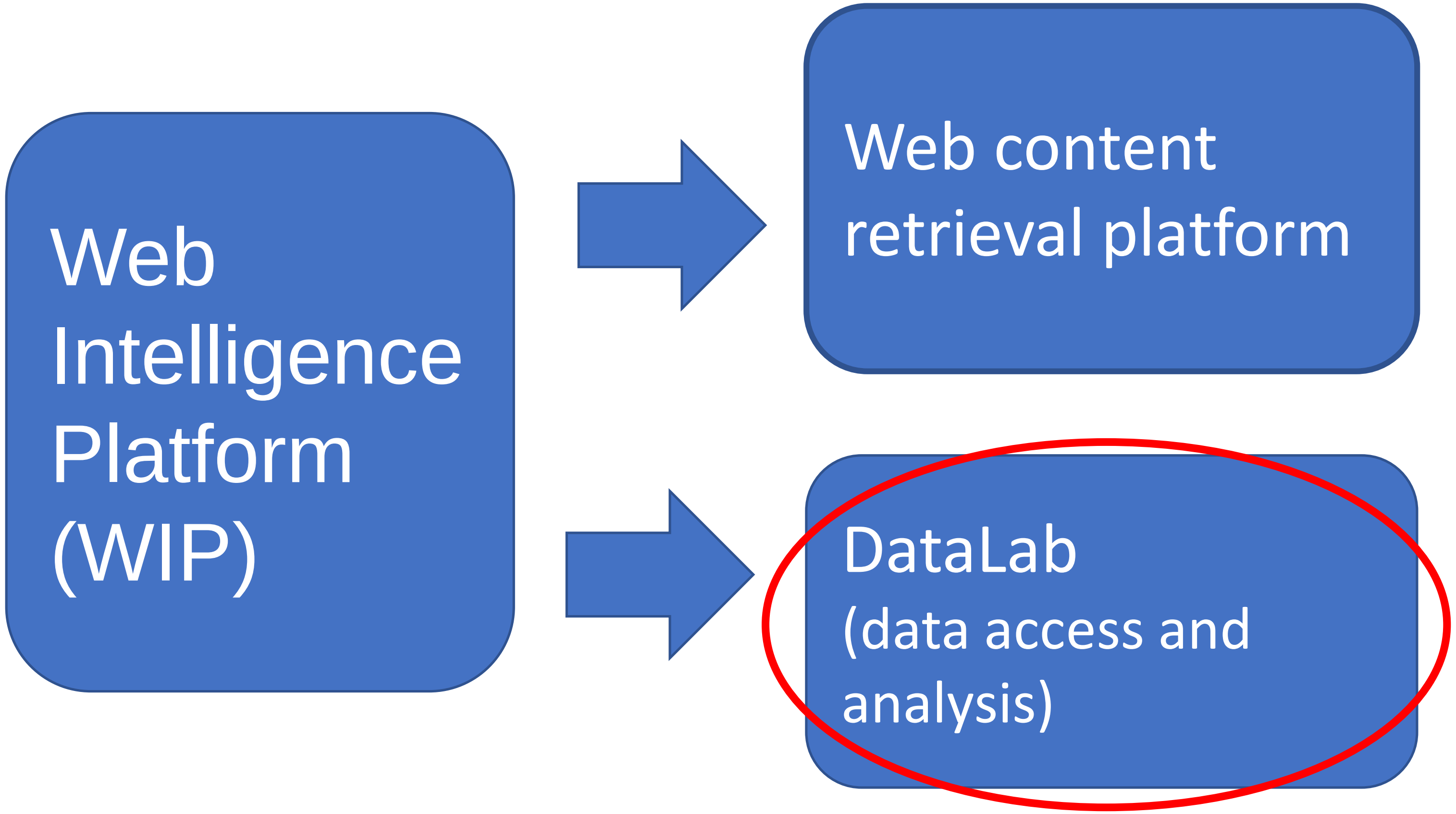
- Job portals for each country selected, based on landscaping process with the involvement of national experts – coordinated by Eurostat/Cedefop

- Goal:

– augment labor market statistics
– provide information on skills

Funded by the European Union

What WIH concept has materialized into to date



Web Intelligence Platform (WIP)

Web content retrieval platform

DataLab
(data access and analysis)

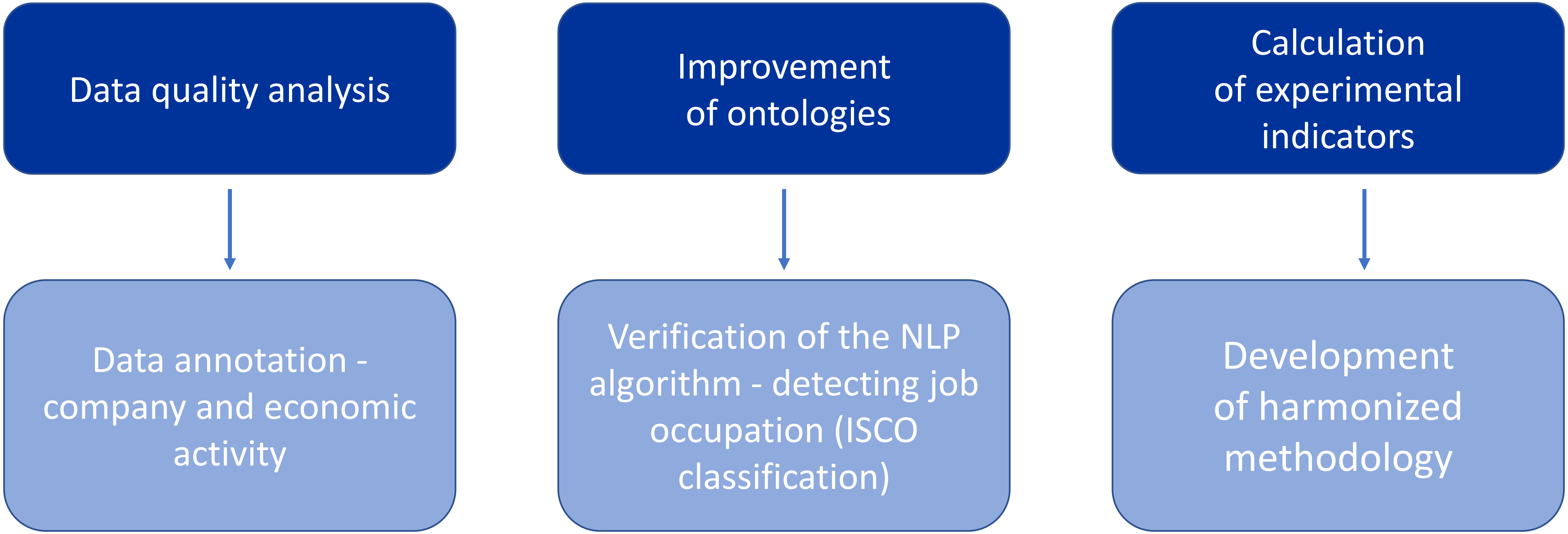
Funded by the European Union

Online job advertisements – key issues for the WIN

- Data source stability
- Quality of the data (overcoverage, undercoverage)
- Quality of data classification (e.g. ISCO, NACE, NUTS)
- Relevance of existing classifications (e.g. ISCO, ESCO)

Funded by the European Union

Online Job Advertisements – what WIN works on



Data quality analysis

Data annotation - company and economic activity

Improvement of ontologies

Verification of the NLP algorithm - detecting job occupation (ISCO classification)

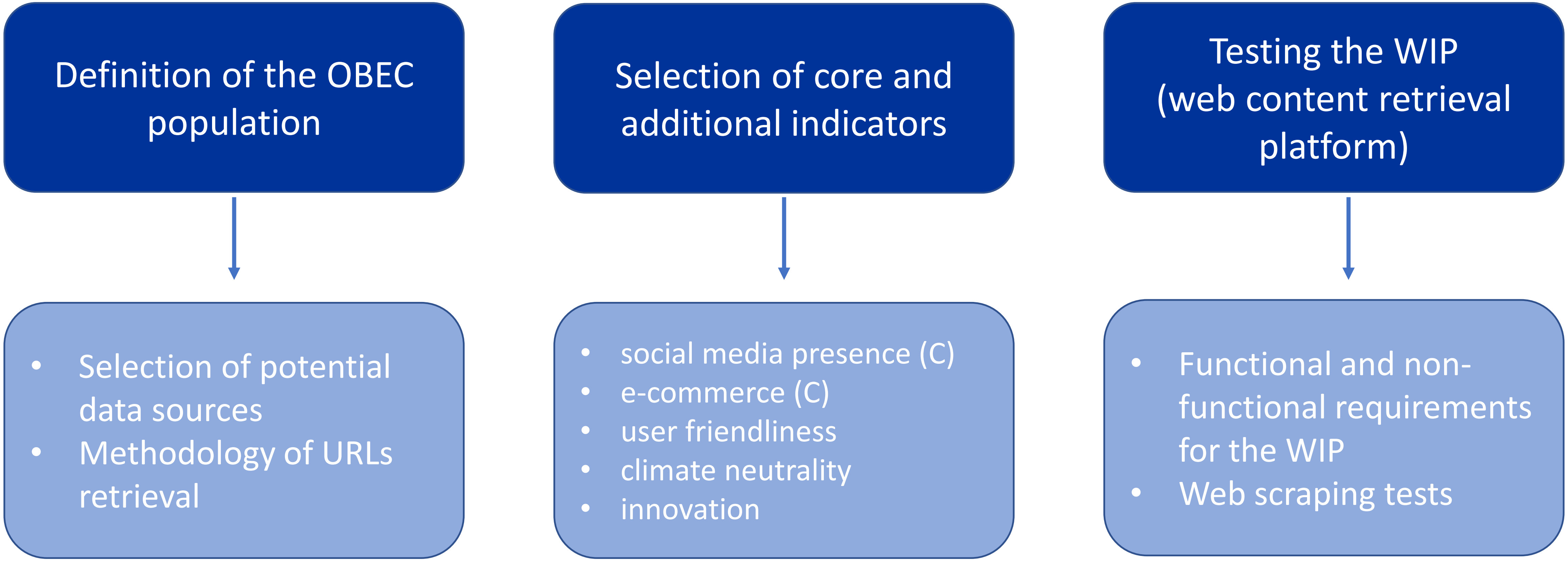
Calculation of experimental indicators

Development of harmonized methodology

Web Intelligence Network

Funded by the European Union

Enterprise Characteristics (OBEC) – what WIN works on



Definition of the OBEC population

- Selection of potential data sources
- Methodology of URLs retrieval

Selection of core and additional indicators

- social media presence (C)
- e-commerce (C)
- user friendliness
- climate neutrality
- innovation

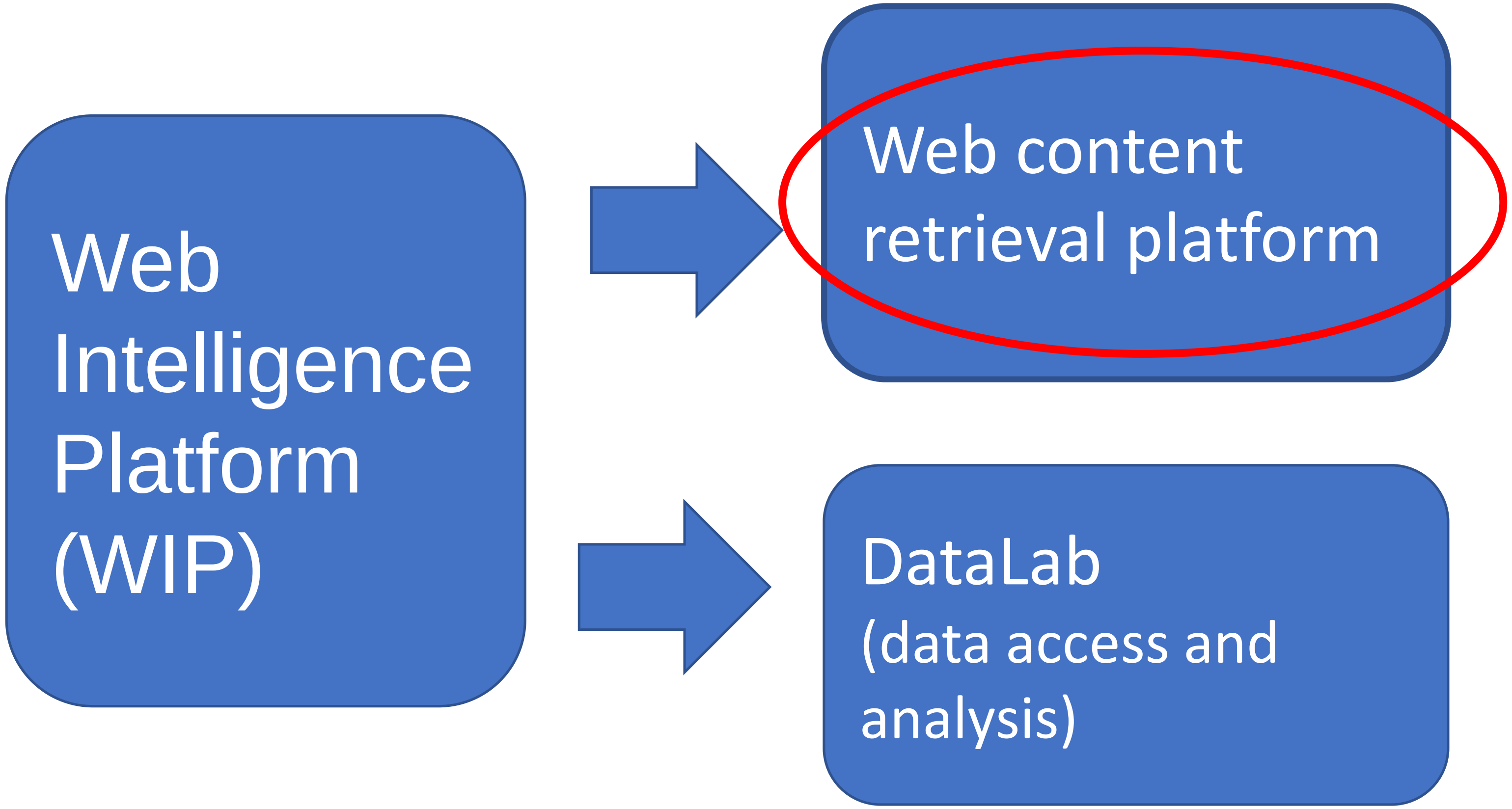
Testing the WIP (web content retrieval platform)

- Functional and non-functional requirements for the WIP
- Web scraping tests

Web Intelligence
Network

Funded by
the European Union

What WIH concept has materialized into to date



Web Intelligence Platform (WIP)

Web content retrieval platform

DataLab (data access and analysis)

Web Intelligence Network

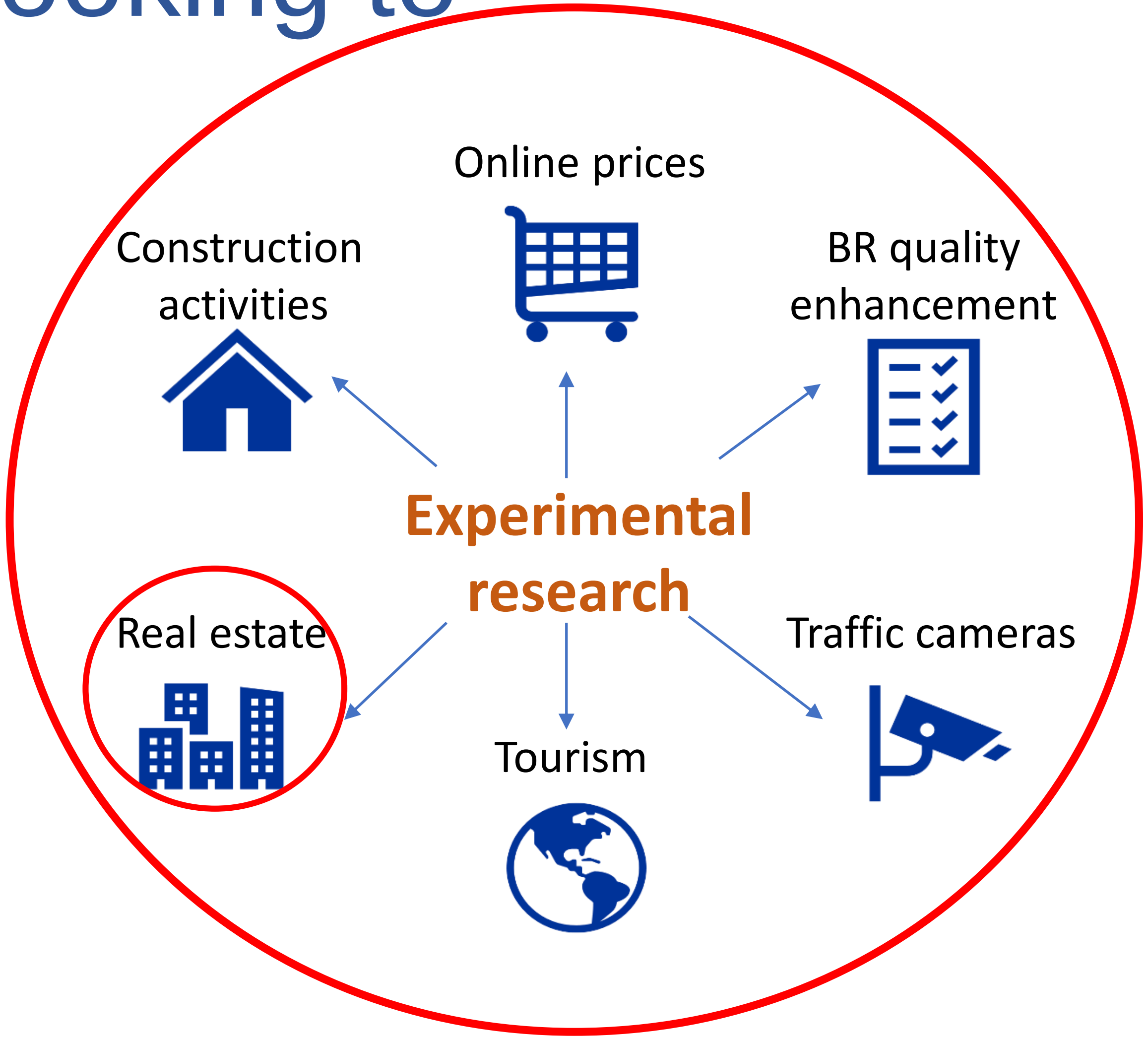
Funded by the European Union

What topics WIN is looking to

Most mature use-cases

Online job advertisements
OJA

Enterprise characteristics
OBEC



Web Intelligence Network

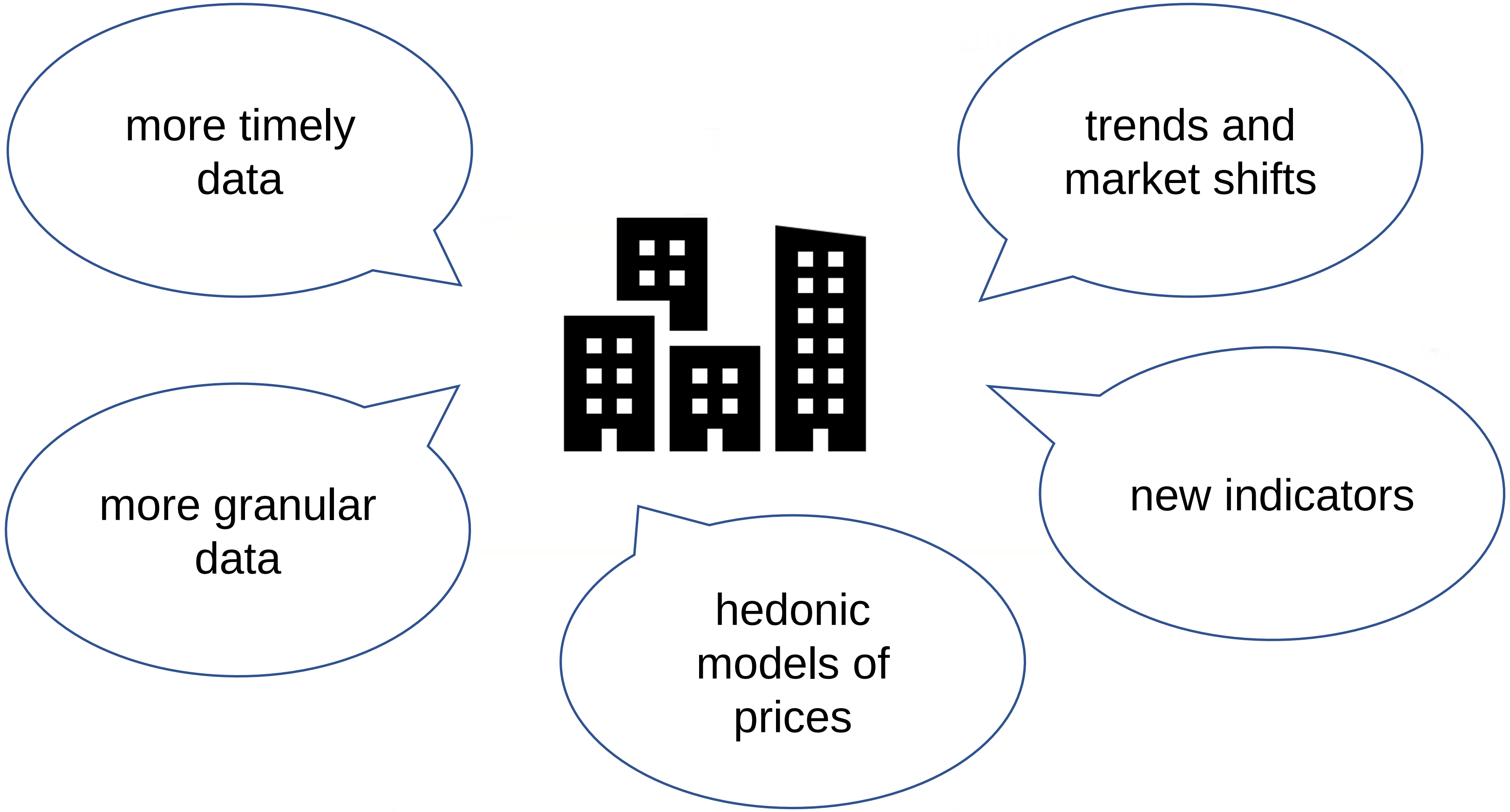
Funded by the European Union

Experimental research – how WIN works

- New data sources exploration & landscaping
- Programming, production of software
- Data acquisition and recording
- Data processing (e.g. de-duplication)
- Modelling and interpretation (i.e. data analysis and quality assessment)
- Dissemination of the experimental statistics and results

Joint work of 2-6 NSIs on each use-case → Solutions applicable at the ESS level

Experimental research – example: real estate



more timely data

trends and market shifts

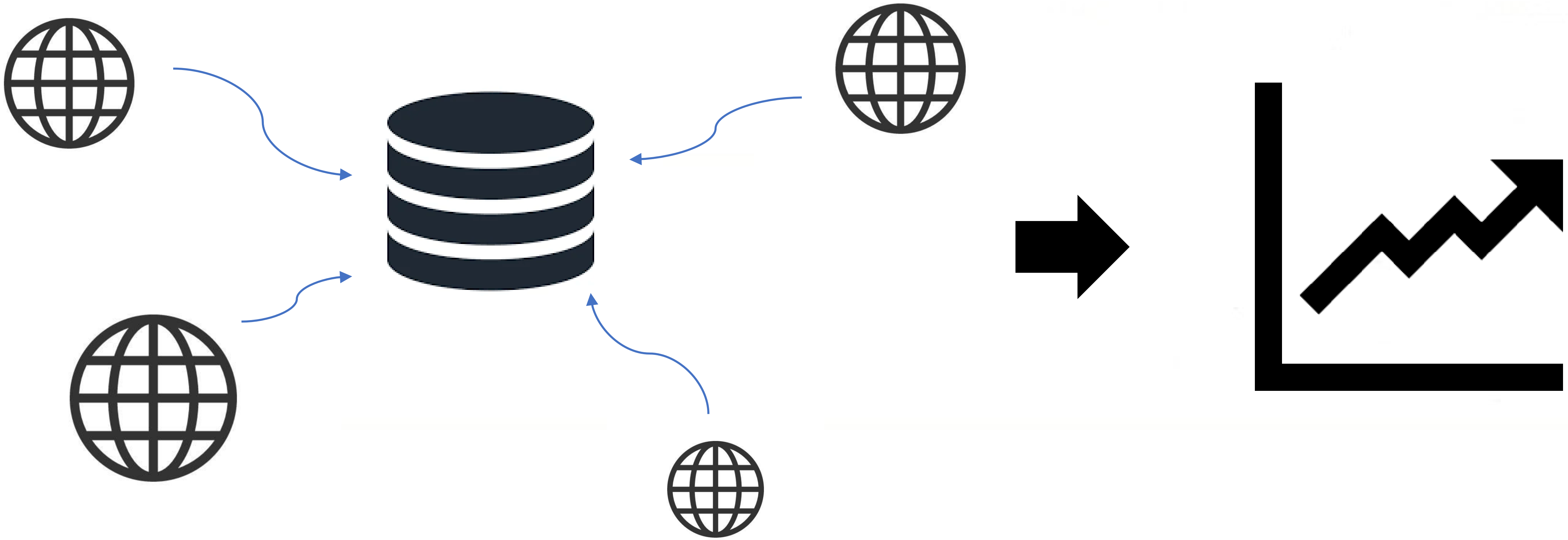
more granular data

new indicators

hedonic models of prices

Funded by the European Union

Experimental research – rental market in Poland



Funded by
the European Union

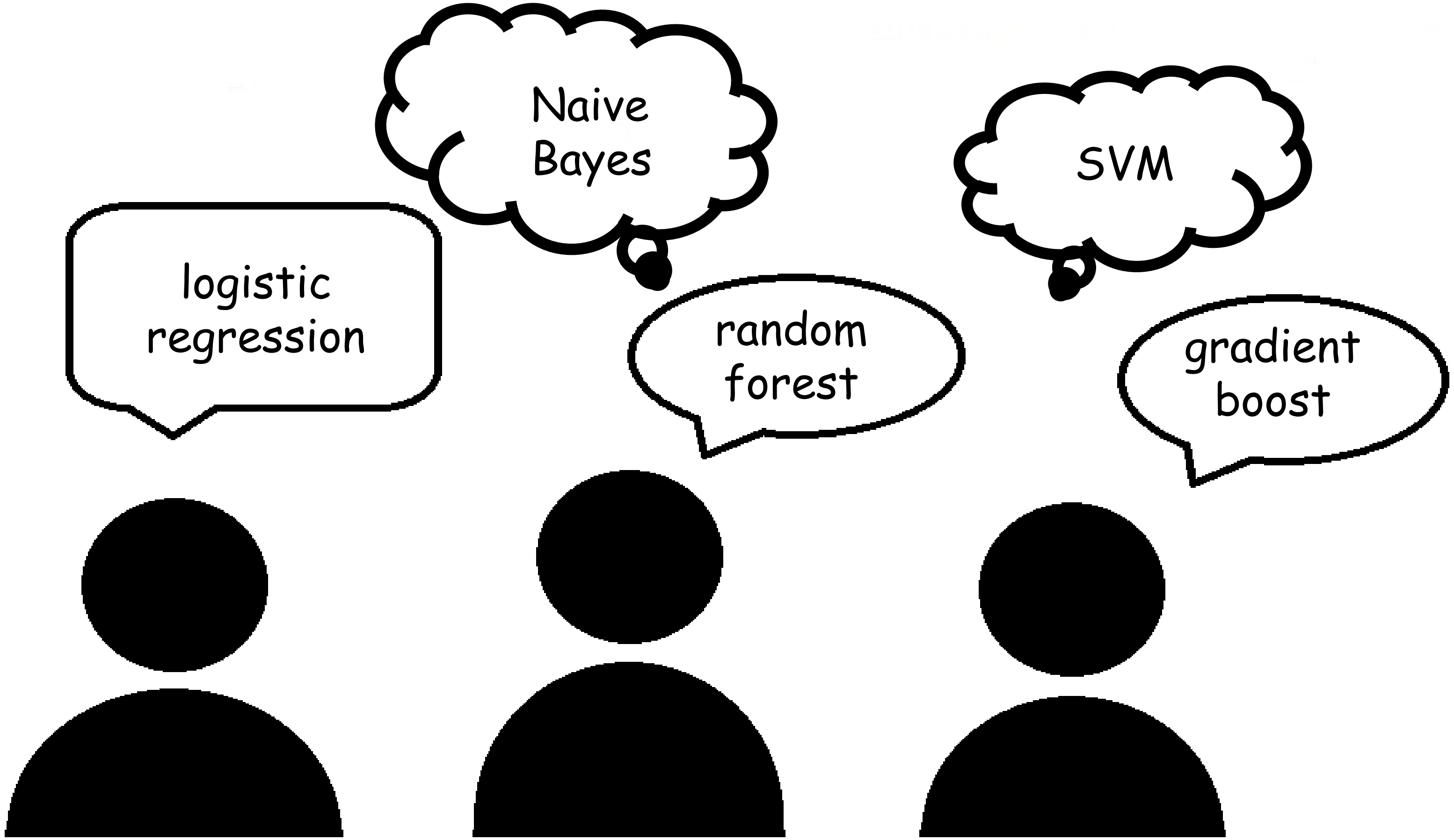
Apartment for rent, Krakow, Poland

We are pleased to present a property with an area of less than 70m2, located in close proximity to the Main Market Square in Krakow. The apartment is located in a historic tenement house, from which the view spreads over the tower of St. Mary's Basilica.

The apartment is located on the 4th floor of a historic tenement house located at the Main Market Square in Krakow. The apartment with a cadastrial area of 67.34 m2 consists of a large living room, 2 bathrooms with a shower and a bathtub, a kitchenette, two separate bedrooms and a spacious mezzanine with a second bedroom. The apartment is fully equipped and fully furnished. From the windows of the premises we can see the charming roofs of the Old Town and the tower of St. Mary's Basilica. The premises is suitable for short-term rental activities and is leased until 2028. With the current layout, it will comfortably provide accommodation for 8 people. The apartment is equipped with air conditioning and video security system.

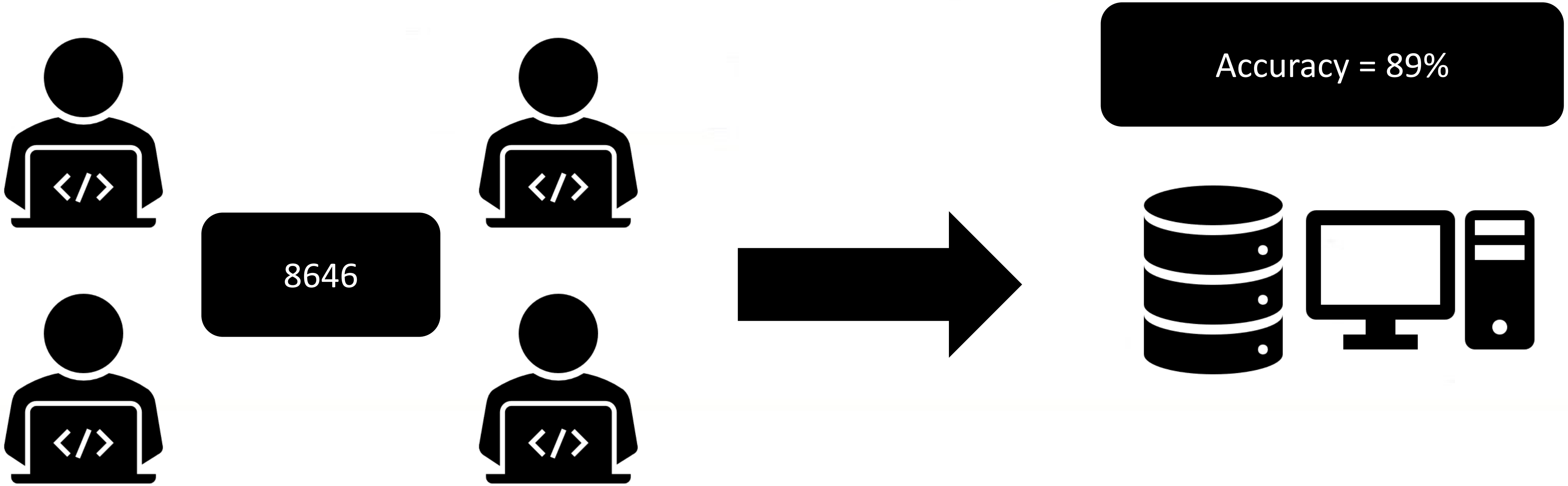
Funded by the European Union

Experimental research – method selection



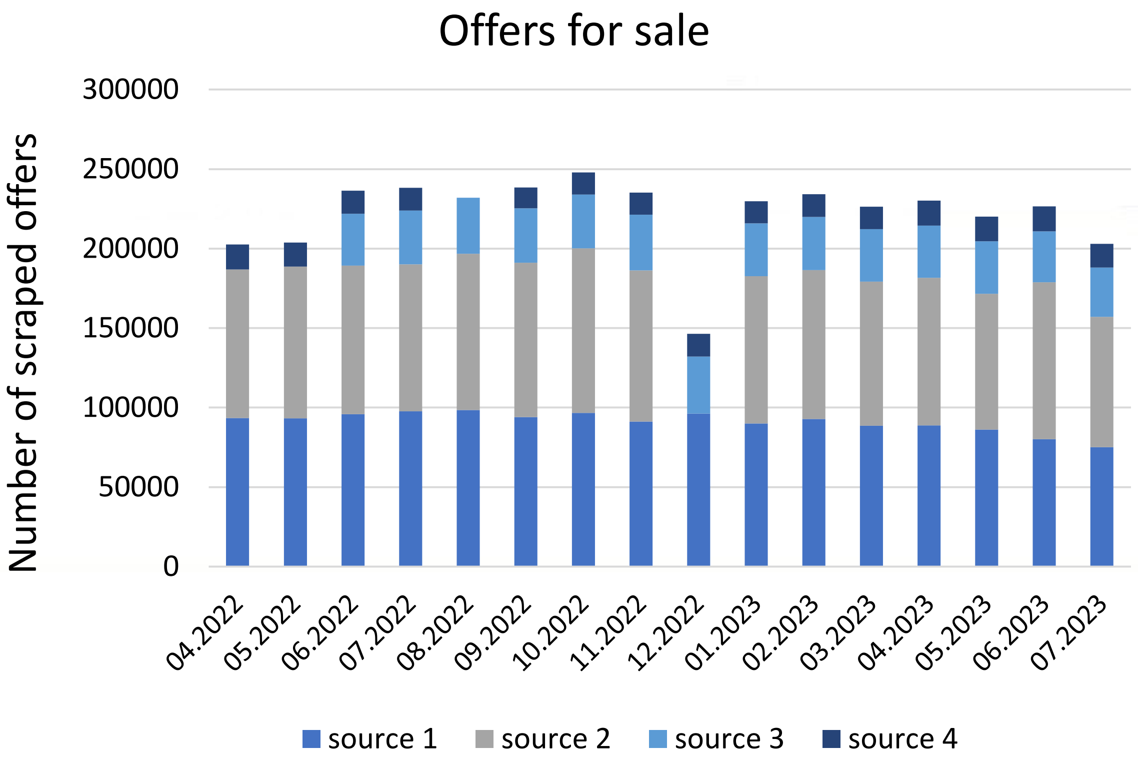
Funded by the European Union

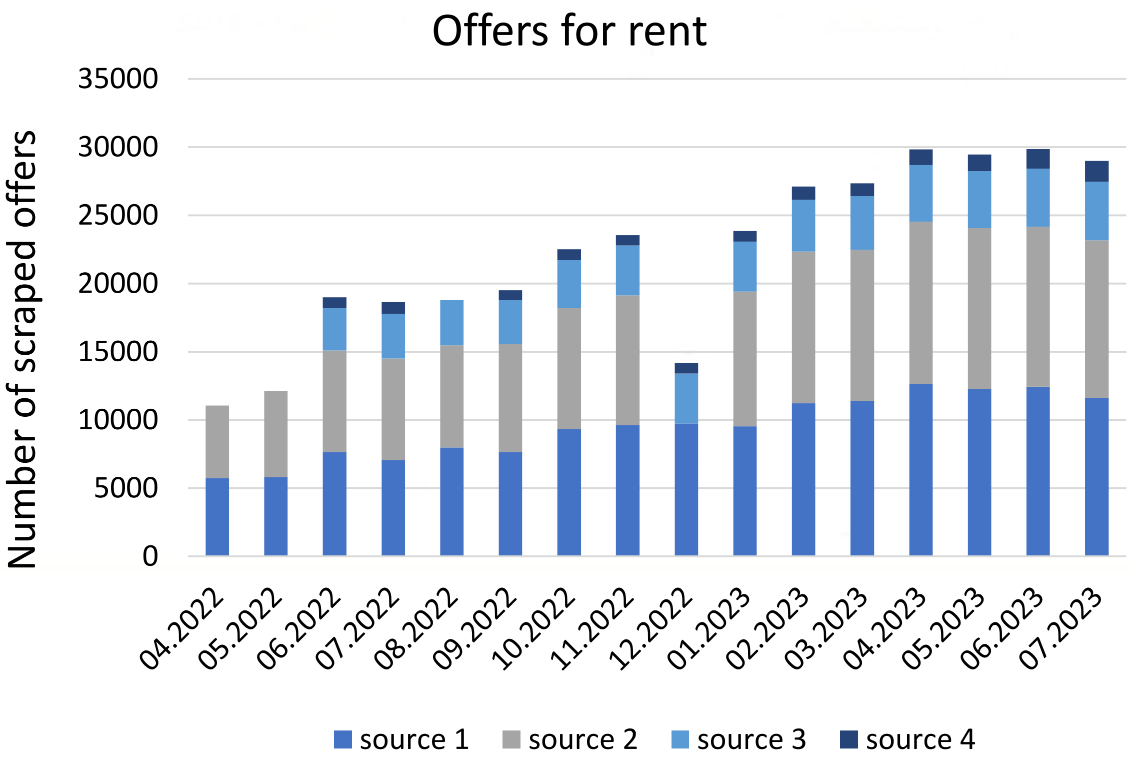
Experimental research – automatic classification of the apartment type



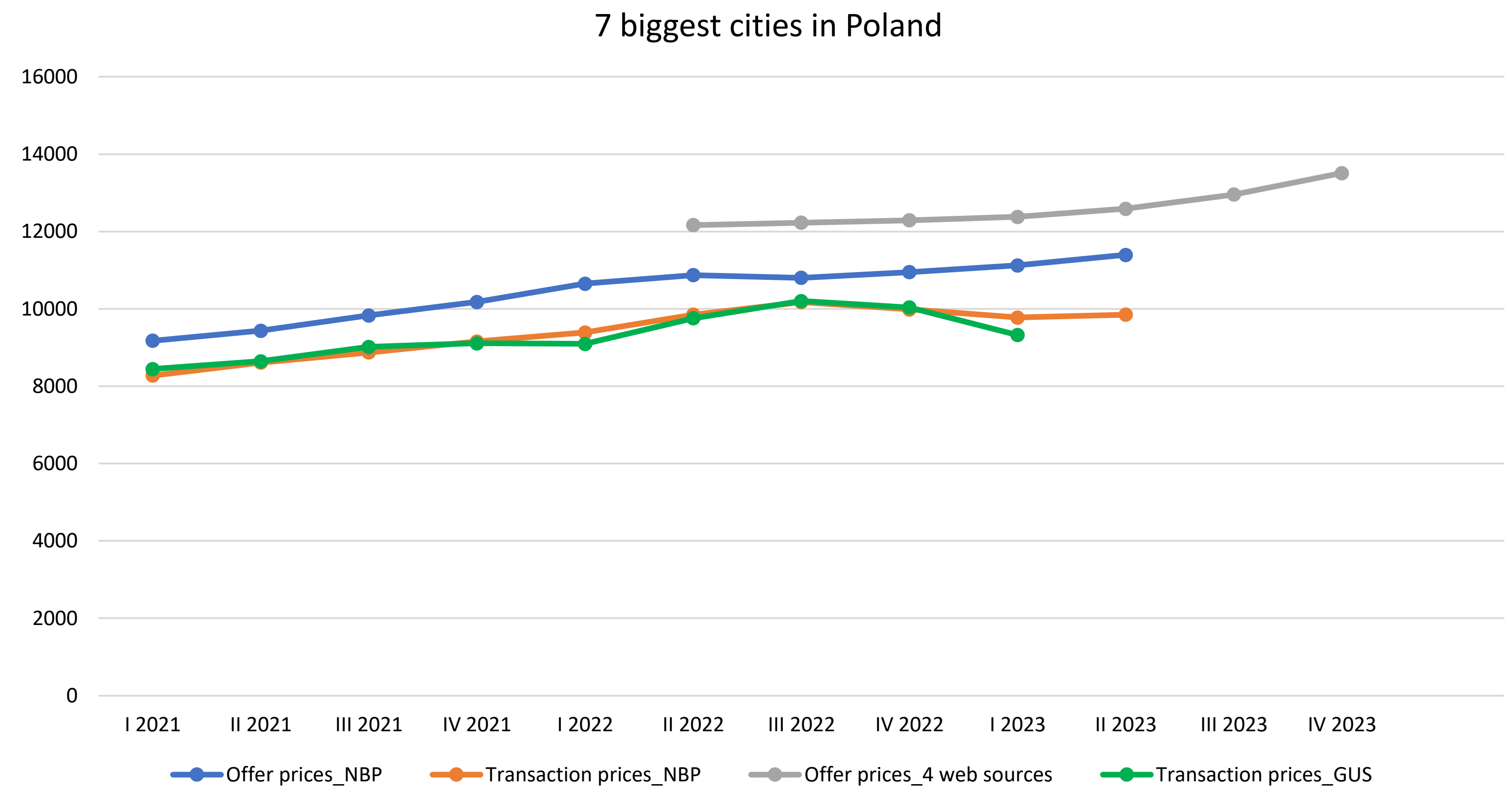
Funded by the European Union

Experimental research – sales /rental offers in Poland





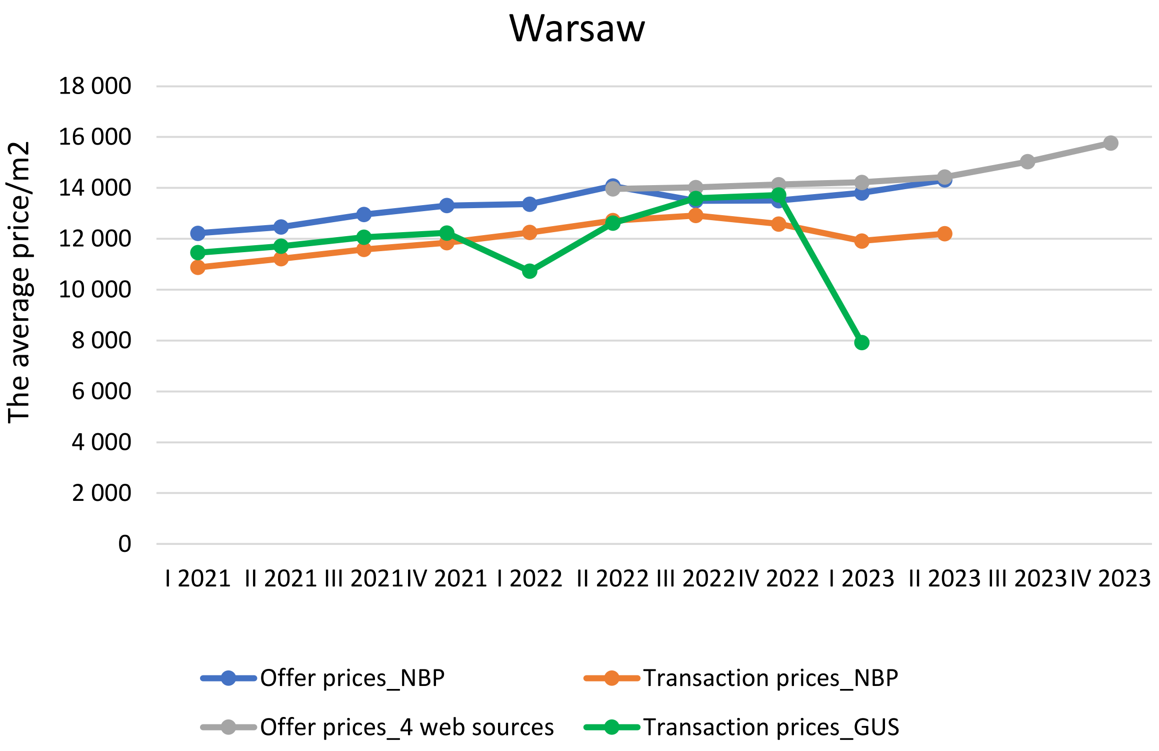
Experimental research - comparison between web data and official statistics

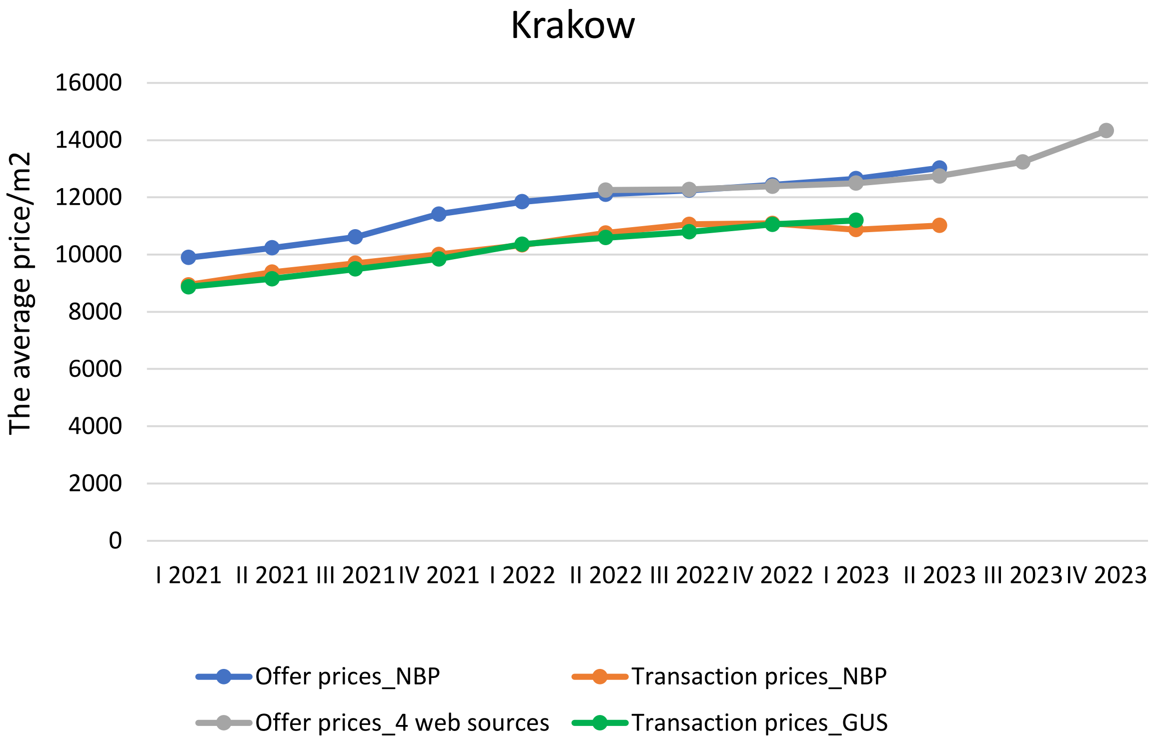


Web Intelligence Network

Funded by the European Union

Experimental research - comparison between web data and official statistics





Web Intelligence Network

Funded by the European Union

Experimental research - what is ahead of us

- Quality assessment and improvement (de-duplication of offers, dealing with missing data etc.)
- Tests of different ML models
- Cooperation with domain experts and validation of results
- Calculation of experimental statistics

Funded by the European Union

Experimental research – quality and methodology issues

- Difference between offer and transaction prices
- Undefined population (lack of reference source for rental market)
- Duplication of offers (within a single portal/ across different portals)
- Multi-offers (apartments in new constructions)
- Missing values (e.g. price of apartments in new constructions)

Funded by the European Union

Challenges - use of web data & WIH by NSIs

Where does the innovation begin?

Who is the „owner” of the development?

Where are domain experts in this process?

Is web data of sufficient quality for official statistics?

Are we ready to sacrifice quality for timeliness?

Do we still need different methods to embrace new data sources for official statistics?

Is our approach to web data acquisition optimal?

Are we facing a shift of the paradigm of official statistics?

Funded by the European Union

We need you to discuss it all!

Meet us at conferences

NTTS conferences

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

IAOS
INTERNATIONAL ASSOCIATION FOR OFFICIAL STATISTICS

Visit us on-line

CROS

Look for our training, webinars, tutorials

- Statistical Business Registers
- Tourism data
- Online real estate market
- Architecture, methodology and quality
- Construction activity
- Web Intelligence in Practice - OBEC
- OJA Training for WIN and WISER
- And more…

Join WISER!!!
Web Intelligence uSers Group

Web Intelligence Network

Funded by the European Union

Thank you…
(…and do not hesitate to contact us!)

do.nowak@stat.gov.pl – Dominika Nowak, ESSnet WIN project coordinator
righi@istat.it - Alessandra Righi, WISER group coordinator

Disclaimer: the presentation uses some of the slides by Eurostat and information developed by the ESSnet WIN colleagues – under their verbal consent

Web Intelligence
Network

Funded by
the European Union