Text mining to identify skills, stakeholders and capabilities: the case of Artificial Intelligence in Emilia-Romagna

A. Quinquilla, N. Duran, F. Massucci, E. Fuster, B. Rondelli (SIRIS Academic)

Leda Bologni, Lucia Mazzoni, Giorgio Moretti (Art-ER)

Objectives

The study presents a semantic analysis to map the science, technology and innovation (STI) activities in the field of Artificial Intelligence in the Emilia-Romagna region (Italy). The study aims at fostering **research-to-business** collaborations; detecting existing and potential **capabilities to address challenge-driven innovations**.

We identified which topics are covered, what is the relative regional specialisation, which are the key actors, and which are the internal and external linkages. Finally, we measured how the topics have been evolving in time and how the **specific domain of Artificial intelligence** is applied to **solve societal challenges (SDGs)**.
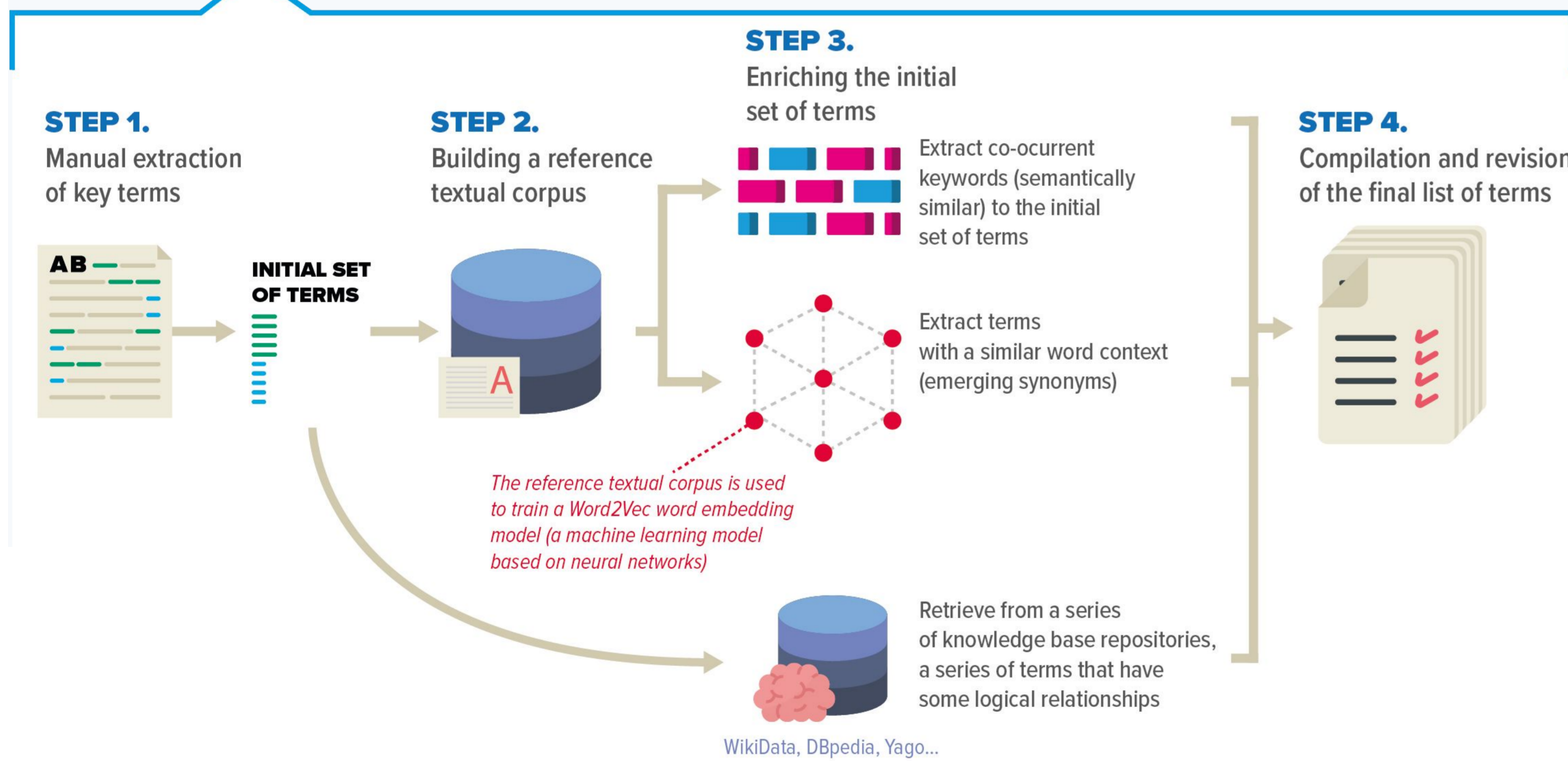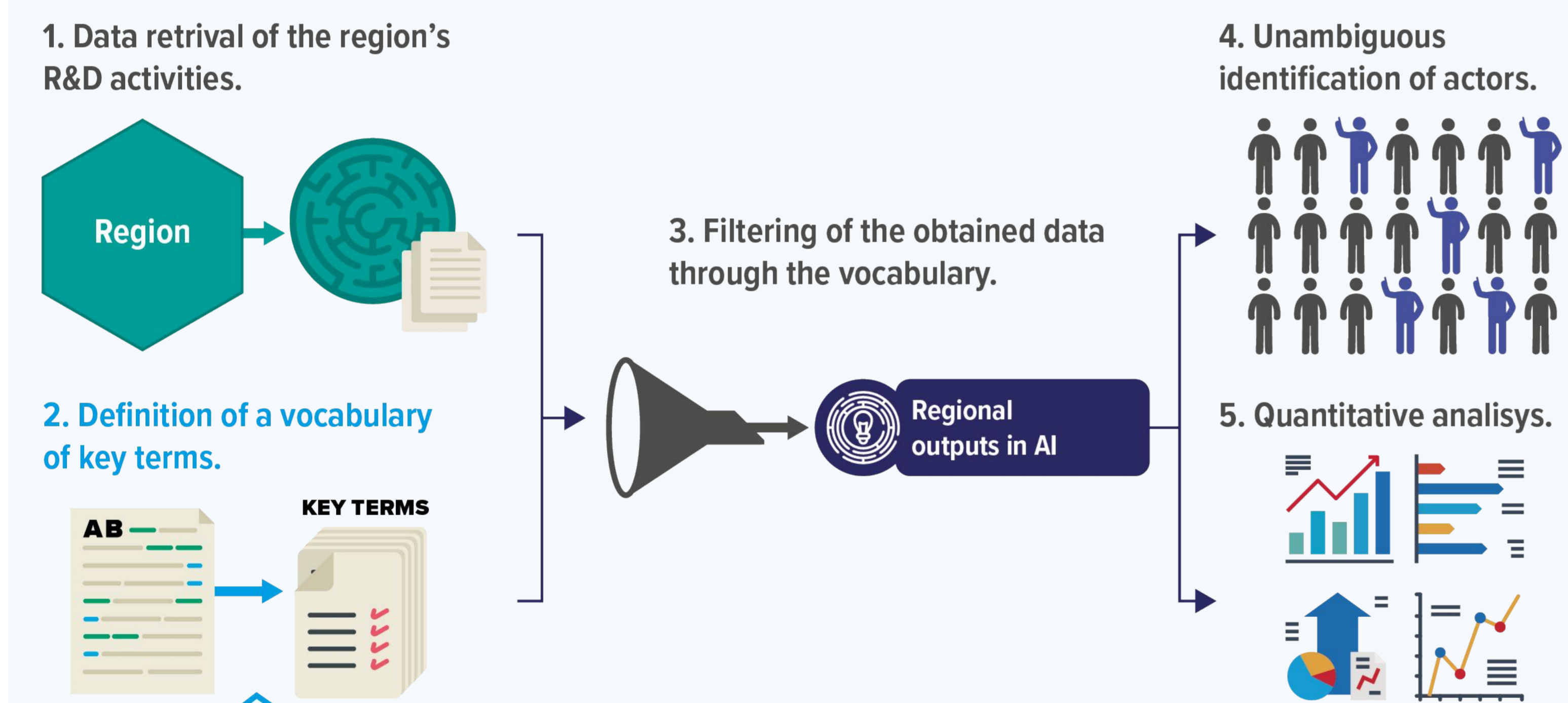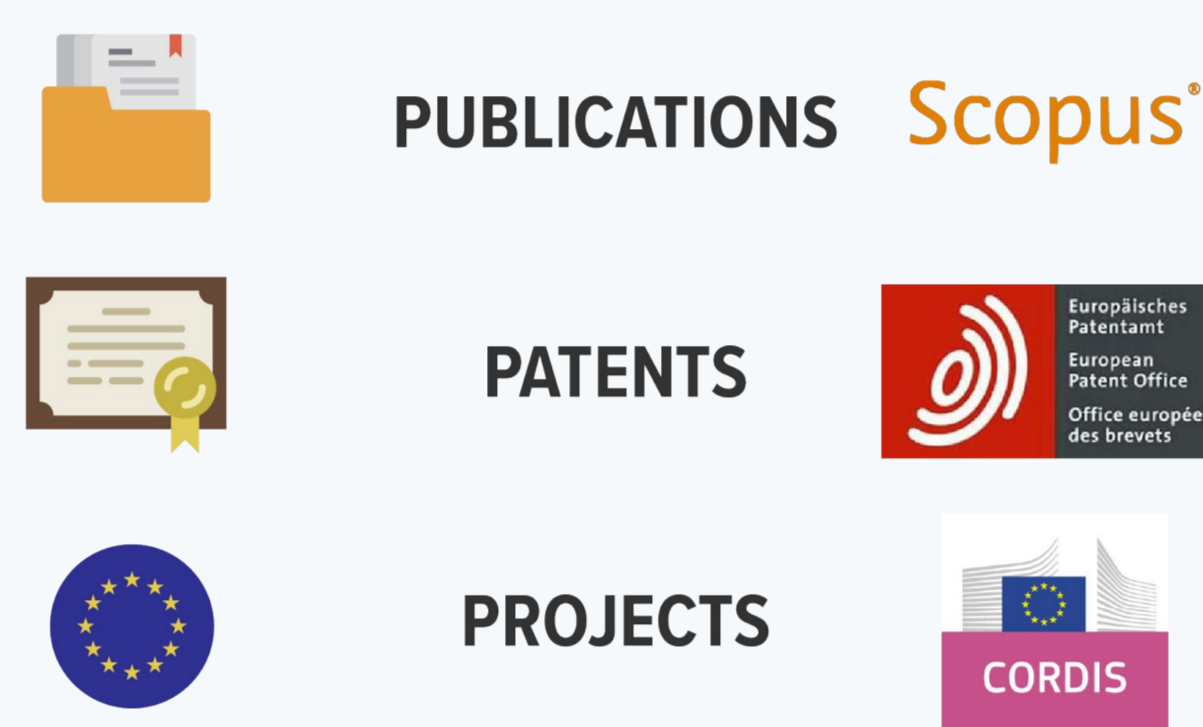
Methodology

We analysed single STI records proceeding from different repositories. Text mining techniques were applied to the publication abstracts, patent descriptions and R&I project objectives to **extract a wealth of textual information describing in detail STI activities and results**.

The methodology uses both automated text-mining computer techniques and qualitative and quantitative analysis.
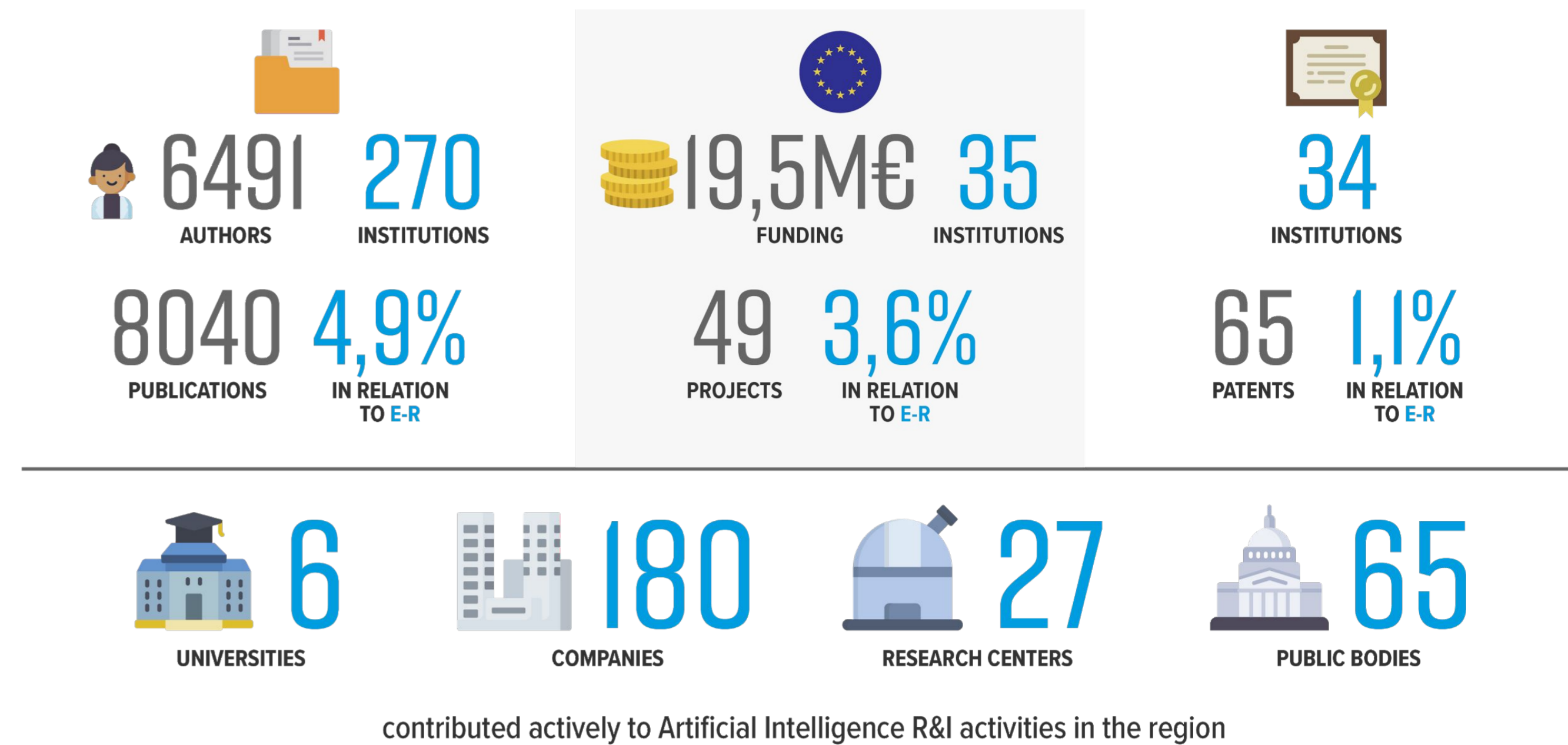
The approach is divided into the following steps:

- Extraction of documents produced by the regional research ecosystem;
- Definition of a vocabulary of key concepts that define the domain of research in Artificial intelligence
- Use of the vocabulary to identify, among the texts extracted in step 1, all those related to the domain of interest
- Fine-grained, unambiguous identification of the actors
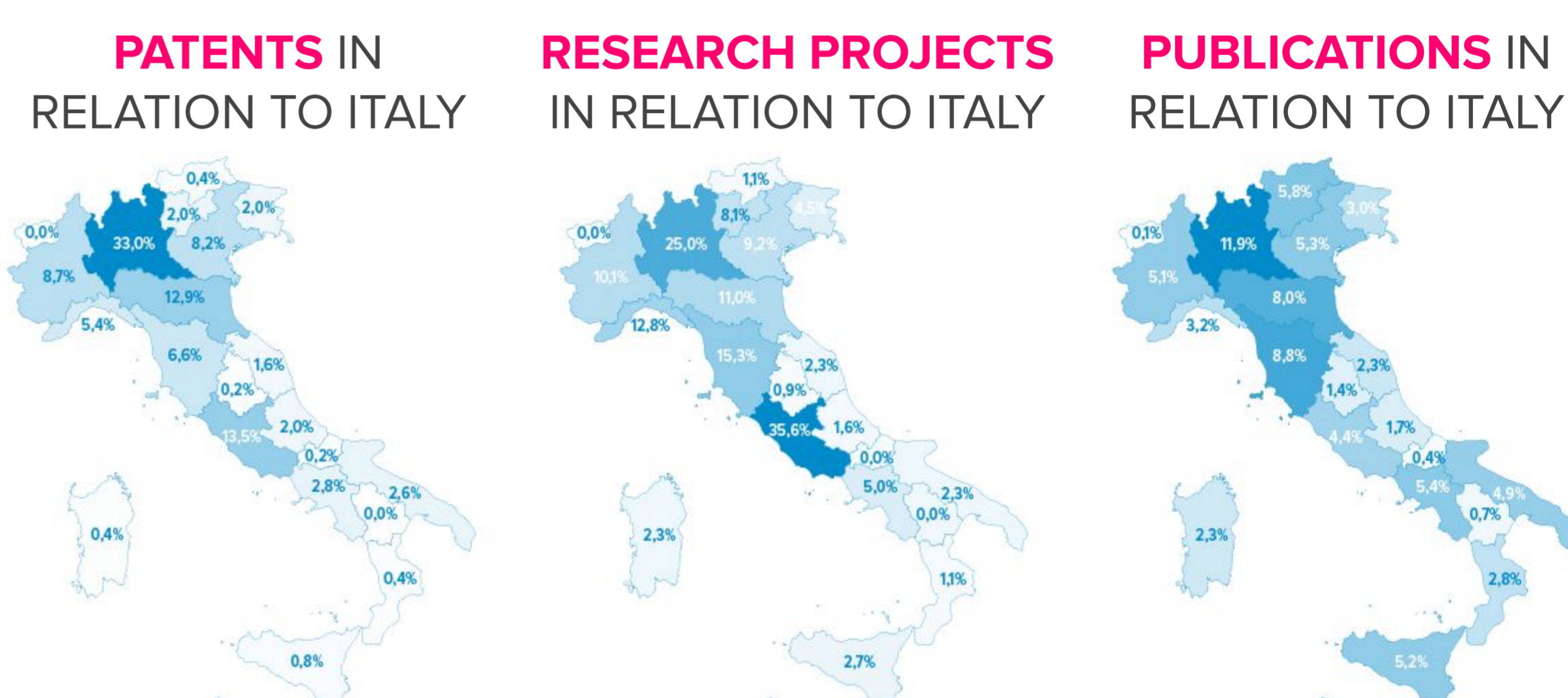- Quantitative analyses using the extracted data

DATA SOURCES

PUBLICATIONS — Scopus

PATENTS — Europäisches Patentamt / European Patent Office / Office européen des brevets

PROJECTS — CORDIS

1. Data retrieval of the region's R&D activities.

Region

2. Definition of a vocabulary of key terms.

AB — KEY TERMS

3. Filtering of the obtained data through the vocabulary.

Regional outputs in AI

4. Unambiguous identification of actors.

5. Quantitative analisys.

STEP 1.
Manual extraction of key terms

AB — INITIAL SET OF TERMS

STEP 2.
Building a reference textual corpus

A

The reference textual corpus is used to train a Word2Vec word embedding model (a machine learning model based on neural networks)

STEP 3.
Enriching the initial set of terms

Extract co-ocurrent keywords (semantically similar) to the initial set of terms

Extract terms with a similar word context (emerging synonyms)

Retrieve from a series of knowledge base repositories, a series of terms that have some logical relationships

WikiData, DBpedia, Yago...

STEP 4.
Compilation and revision of the final list of terms

Results

6491 AUTHORS — 270 INSTITUTIONS

19,5M€ FUNDING — 35 INSTITUTIONS

34 INSTITUTIONS

8040 PUBLICATIONS — 4,9% IN RELATION TO E-R

49 PROJECTS — 3,6% IN RELATION TO E-R

65 PATENTS — 1,1% IN RELATION TO E-R

6 UNIVERSITIES — 180 COMPANIES — 27 RESEARCH CENTERS — 65 PUBLIC BODIES

contributed actively to Artificial Intelligence R&I activities in the region

PATENTS IN RELATION TO ITALY

RESEARCH PROJECTS IN RELATION TO ITALY

PUBLICATIONS IN RELATION TO ITALY

Specialisation analyses: here we see the percentage of patents, European projects and publications, by each Italian region with respect to the total national production.
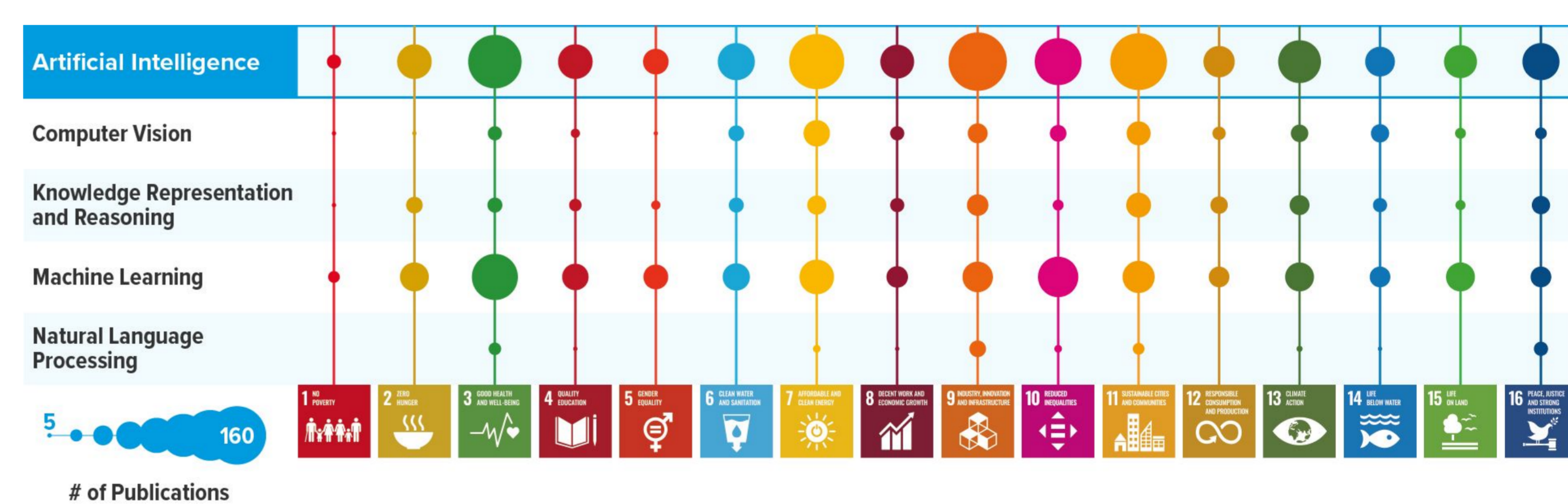
Universities and Research Centers
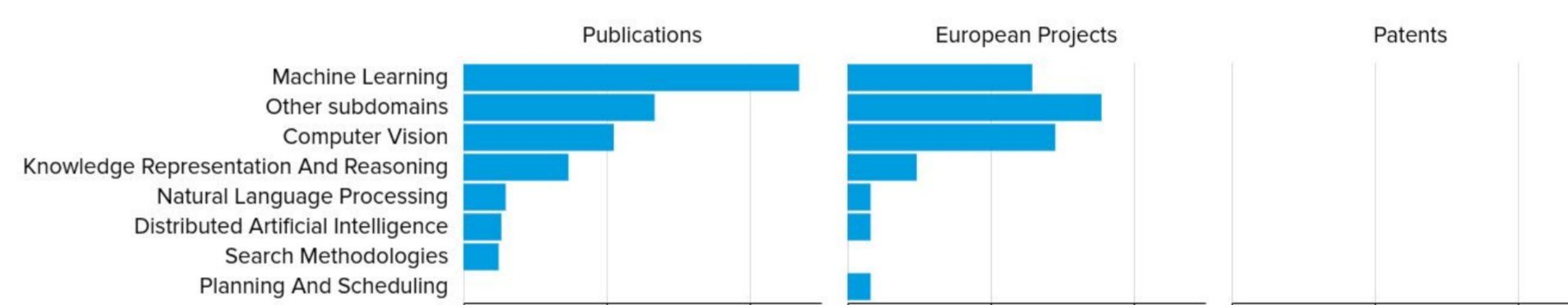
Companies

Percentage of patents, European projects and publications, classified by **sub-fields of AI**, produced different actors.
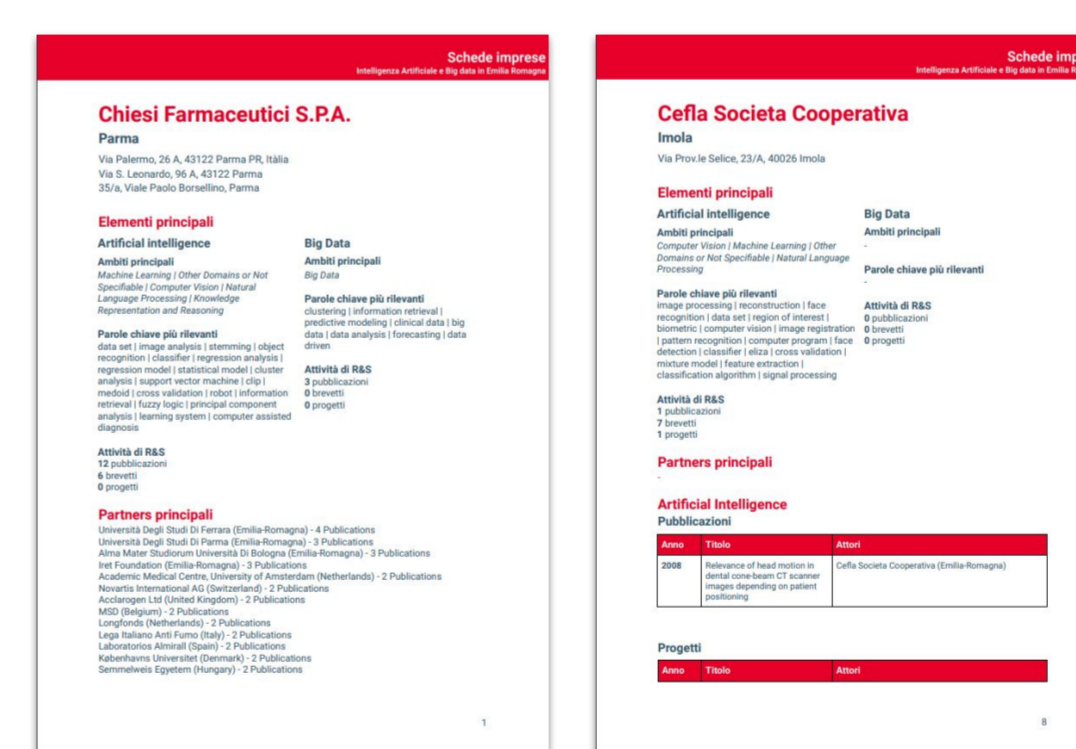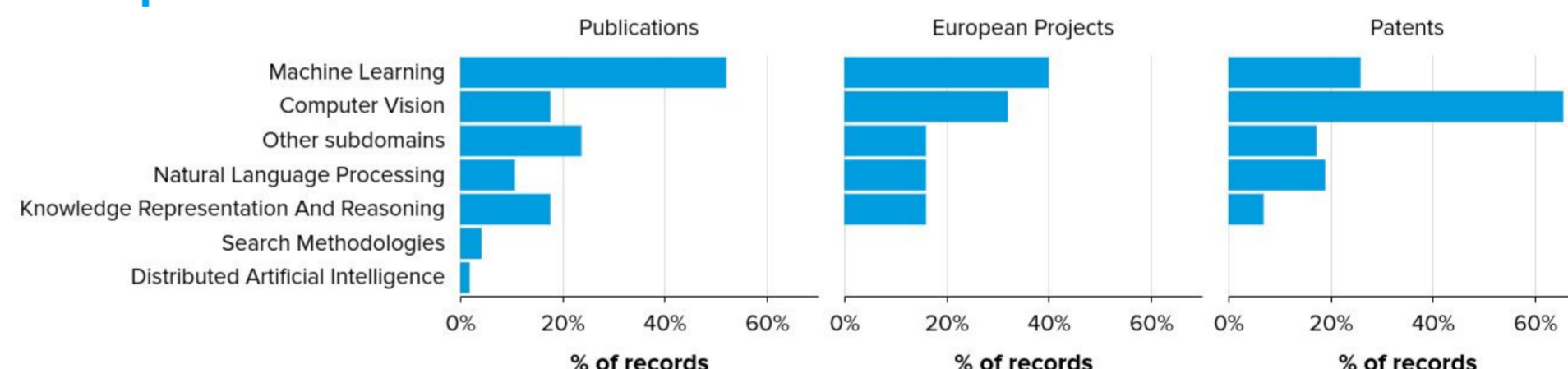
A **profile card** was automatically generated for each **private company**, presenting their main topics of activity in the AI domain and the partnerships, at national and international level.

Artificial Intelligence

Computer Vision

Knowledge Representation and Reasoning

Machine Learning

Natural Language Processing

5 — 160 # of Publications

Outputs in Artificial Intelligence were classified identifying the **contribution to the different SDGs**. Here we see, as example, how many publications of AI (and sub-fields) in Emilia Romagna are contributing to the different SDGs.

Conclusions

The advantage of the proposed approach is that, by being built on a set of controlled terms positively linked with each specific domain and respective subdomains, it allows for a clear cut identification of textual records surely related with specific domains and subdomains. The case **study here proposed is just a possible application of the methodology, which can be replicated for different domains of interest** (e.g. Industry 4.0, Agrifood, etc.), **with useful applications for public decision-makers and private companies**. For the former, our method allows to get, for specific domains, a precise cartography of local actors and pockets of excellence. For the latter, it enables to identify competitors and potential partners in specific R&D niches.