Big Data for Labour Market Intelligence

Capacity development programme 2022

Module 2: Dissemination and analysis

Session 4

The role of AI - ML in the data system, classification, analysis

Speaker: Fabio Mercorio

16/11/2022

Roadmap

1. Changing factors in LM
2. On the contribution of Big Data and AI
3. Real-life examples
4. Why should we trust AI? Towards XAI
5. Concluding Remarks

Something has been changing….

Two main factors responsible of LM changes

Structural changes related to Macro Trends

- **Technological innovation** is reshaping jobs towards the realisation of a digital revolution
- **Population ageing:** 65y+ doubled worldwide between 1990 and 2020
- **Work mobility:** 20% of the European workforce was born in a Country different from the working one
- **Sustainability** is growing within job demand asking for green skills

LM is changing fast in terms of

Jobs

- Digitalisation and the rise of technology is reshaping jobs: several jobs are **disappearing** while novel jobs are **emerging**;
- of these some are simply a variant of existing jobs, others are genuinely new jobs that were non-existent until a few years ago.

Skills

- **Skills and competencies are reshaping jobs even more**: *digital skills*, *soft skills* along with *green skills* are growing in importance in many jobs, even for those not related to technology or the environment as well.
- Skills define the job, and not the vice versa
- **Lifelong learning**: Upskilling/reskilling are even more important

Substitution

- **The impact of technology and robotisation is affecting the labour marketing**, exposing jobs with low level of «thinking skills» and «manual dexterity» to a high probability of being robotised in the next few years, even partially [Frey and Osborne, 2017]

The need for knowledge on labour market

Monitoring occupations and skills…

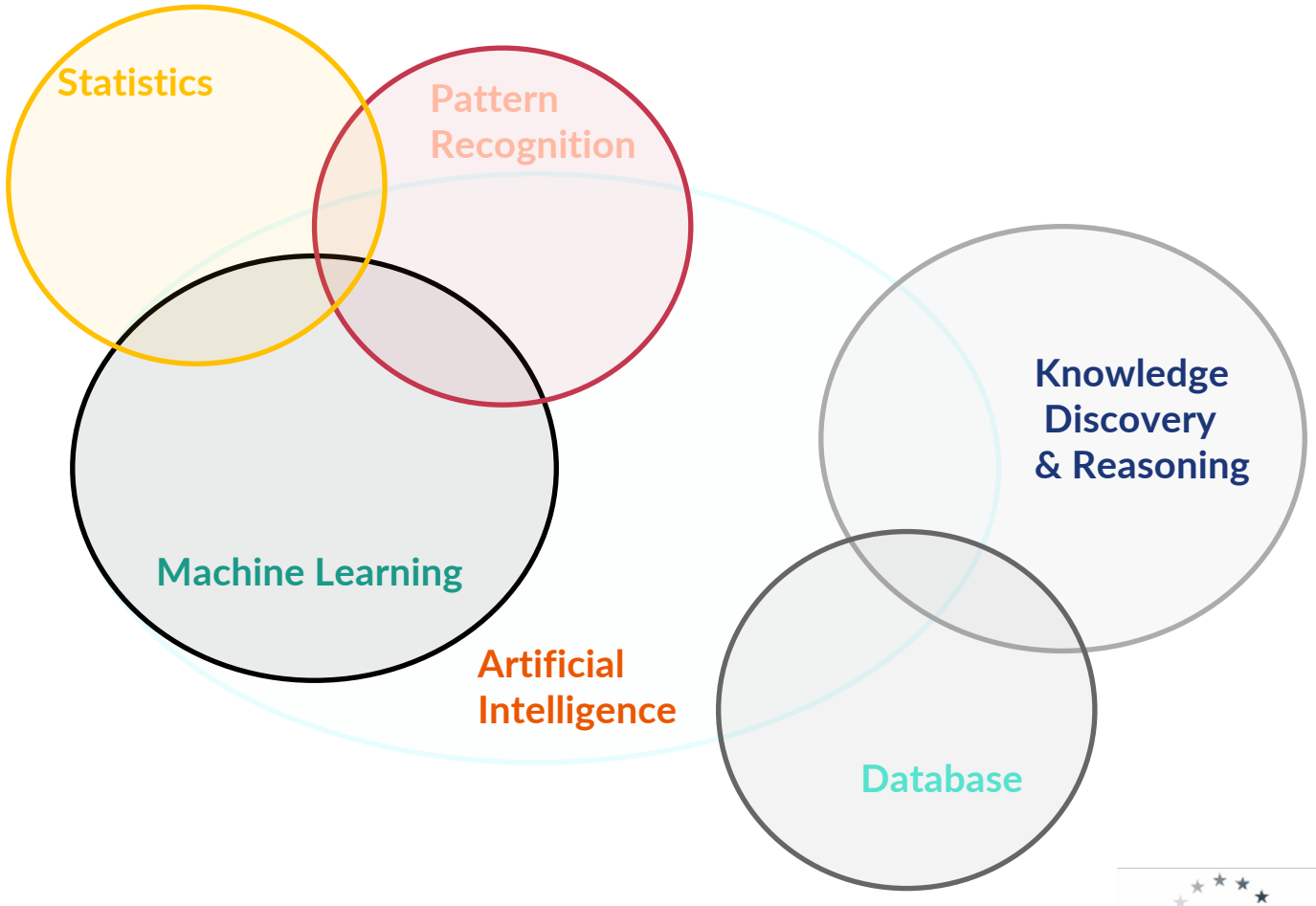
- The need for knowledge to monitor and understand labour market dynamics promptly and to define policies timely becomes key for workers, companies, LM operators and policymakers
- In such a dynamic environment, observing **Skill changes** requested by companies is key to sustaining an inclusive labour market reducing the **skill mismatch**

How AI and Big Data Analytics can contribute?

Big Data is the fuel of AI

Artificial intelligence (AI) refers to **systems** that **show** intelligent behaviour: by analysing their **environment** they can perform various tasks with some degree of autonomy to achieve **specific** goals.

European Commission. AI for Europe (2018)



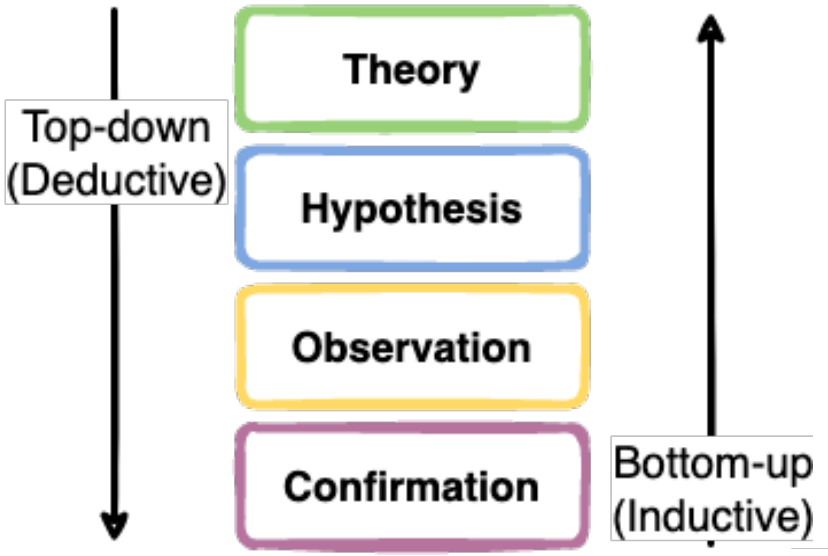
ETF Working together Learning for life
European Training Foundation

A paradigm shift

Big Data: large-scale data, usually having a varied and complex structure. They
Big Data Analytics (BDA): processing big data and looking for valuable information, correlations and patterns within them.

The use of BDA to support decision making requires a paradigm shift (not a replacement!) in the way decisions are made, moving from deductive reasoning to inductive reasoning.

Both deductive and inductive approach are needed!



ETF Working together Learning for life
European Training Foundation

Big data challenges

Data Lifecycle

Data Challenges

Characteristics of data that make handling and managing them challenging

- Volume
- Velocity
- Veracity
- Variety
- Value

Process Challenges

Include all the challenges that make the encountered while getting and processing to the data

- Data selection
- Processing
- Transformation
- Data Analytics

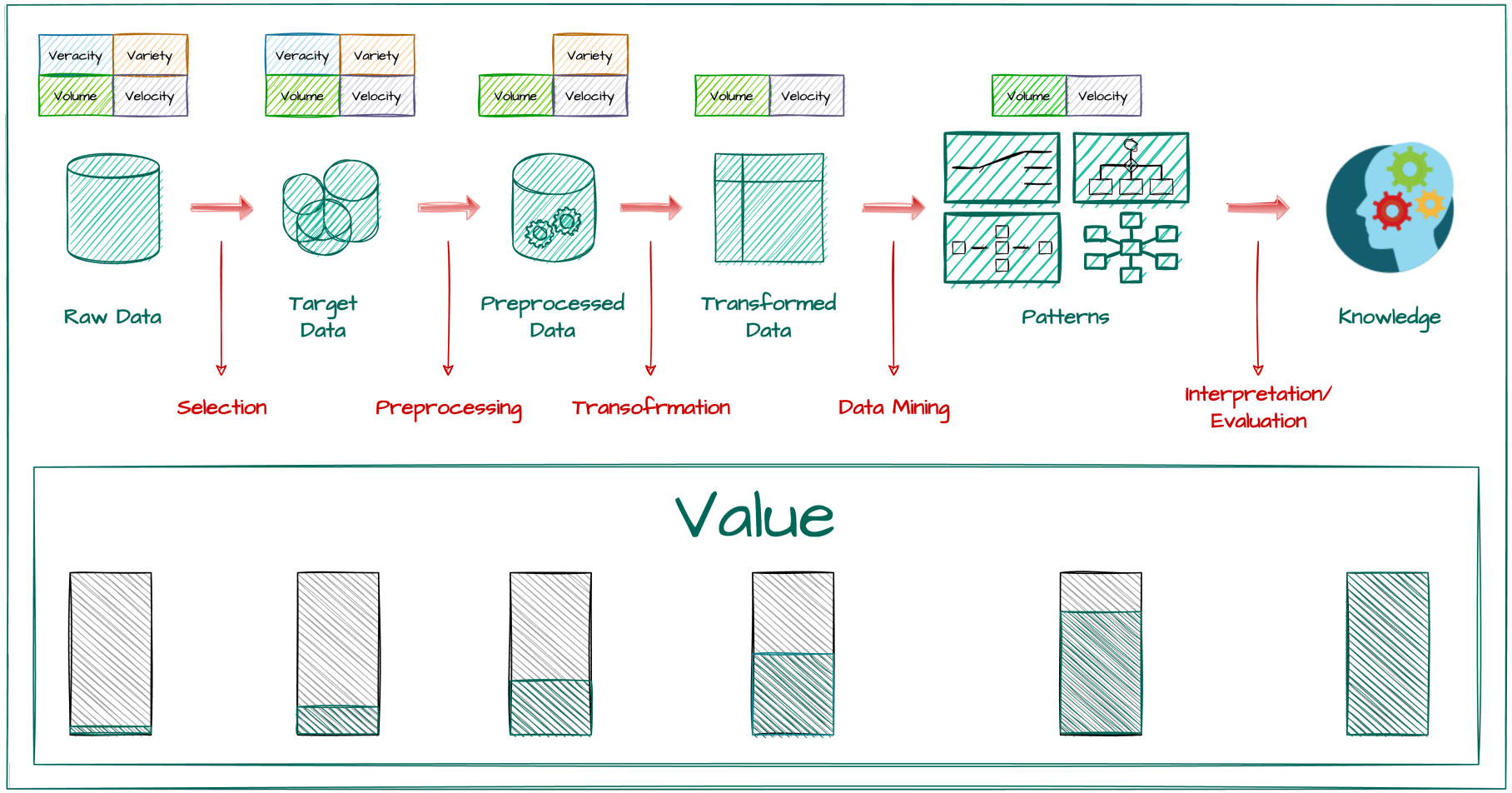
Management Challenges

challenges that an organisation faces while building their decisions on big data

- Data Privacy & Security
- Credibility & Believability
- Data Source Stability
- Representativeness
- Maintenance Costs
- **Transforming Knowledge into business actions**

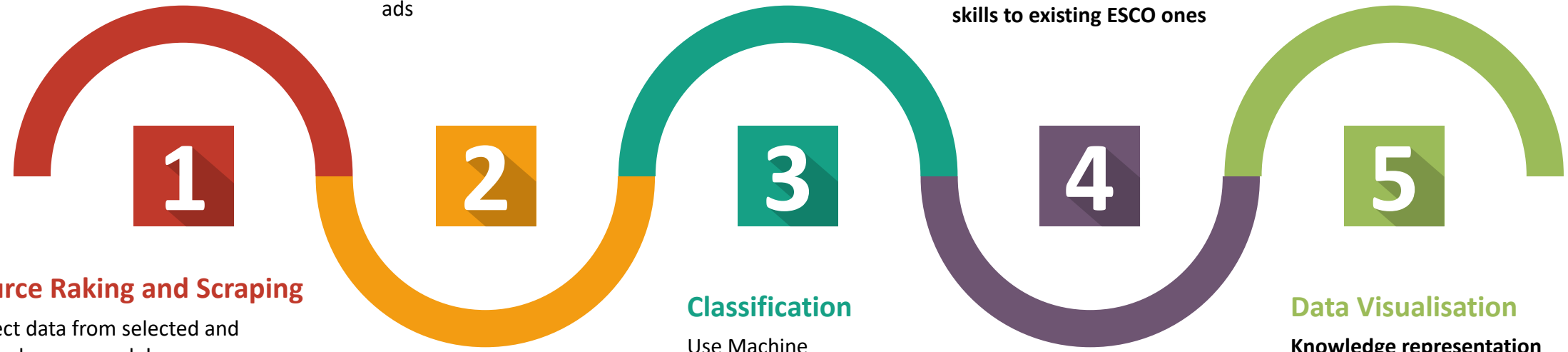
ETF Working together Learning for life
European Training Foundation

The big data lifecycle



ETF Working together Learning for life
European Training Foundation

Transforming LM Information into Knowledge



Source Raking and Scraping

Collect data from selected and trusted sources only!

Data transformation and cleansing

Identify duplicated job ads

Classification

Use Machine Learning approaches to c**lassify Job ads** according to ISCO/ESCO classification systems

Skills extraction

Use top AI algorithms **to extract, recognise and link skills to existing ESCO ones**

Data Visualisation

Knowledge representation of data evolution over time by means of data **visualisation paradigms,** having in mind the stakeholder peculiarities and abilities in reading the data

ETF Working together Learning for life

European Training Foundation

Let’s see some real examples

For Policy makers and labour market experts

Real-life Application: The Index of Change

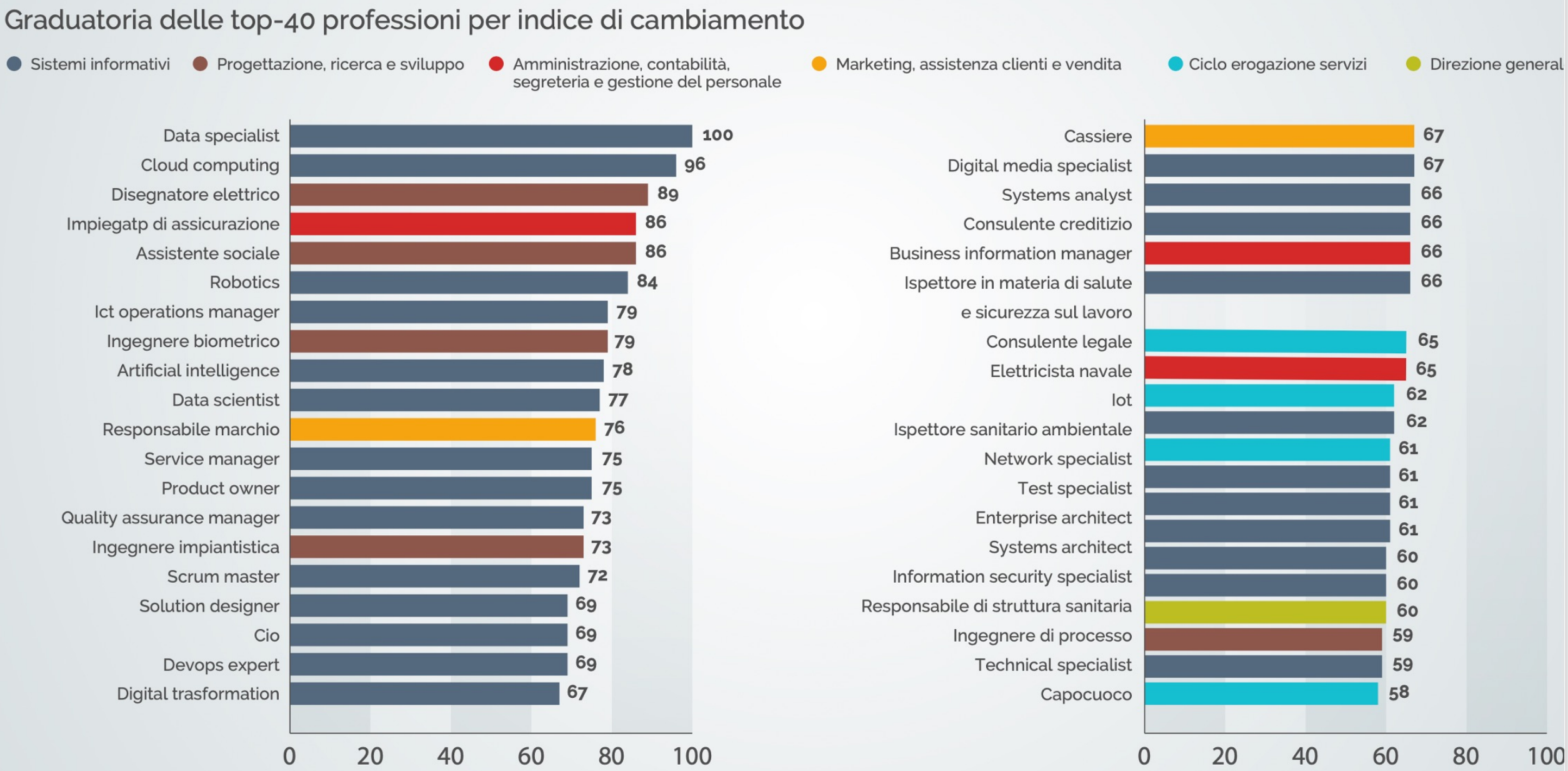
Analysing million Italian Job Ads to identify skill novelty within occupations

Index of Change: how skill demand changes over time

The **Index of changes is built on 3.1 million italian** job ads from 2015 to 2019 with the aim **to support** LM experts for policy evaluation and design by

1. **Understanding Vacancy trend** as the percentage variation of number of vacancies over time
2. **Skill Set Novelty**: identifies the percentage of skills added to the skill set in 2019 compared to 2015. The skill set is the set of competencies common to an area (or a professional figure) over a period of observation.
3. **Variation of relevance in the core set** estimates the relevance of common skills between 2015 and 2019

Index of Change: how skill demand changes over time



For PES

Real-life Application: New Emerging Jobs

Analysing million Job Ads to identify novel skills and jobs

ETF Working together Learning for life
European Training Foundation

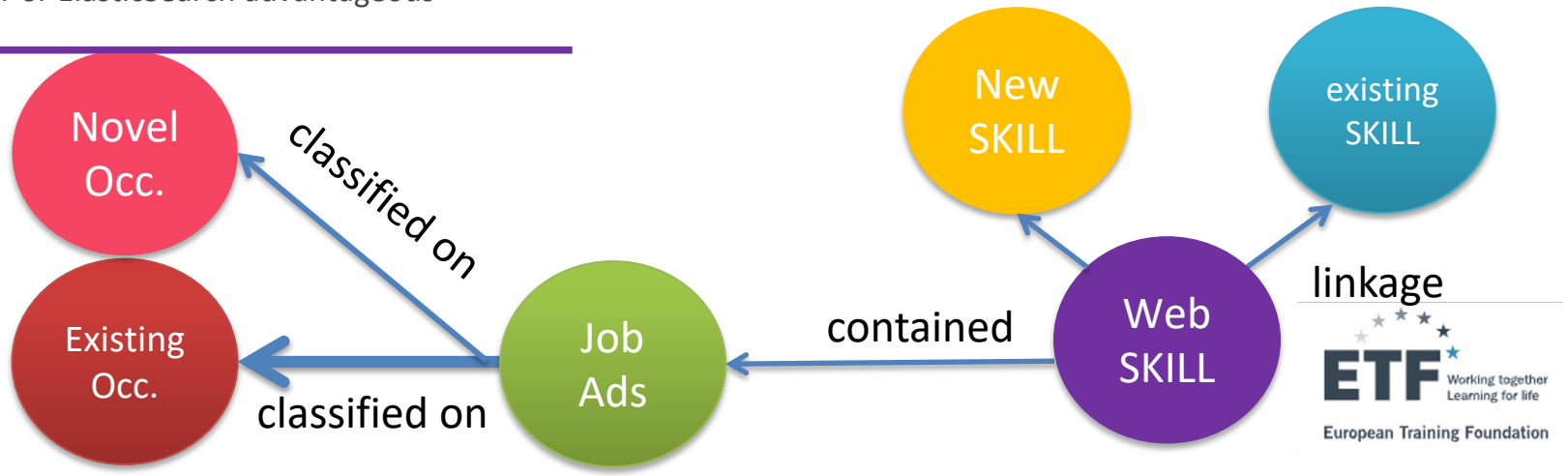
Online Job Vacancy example

Job Title: Data Scientist.

Description: We're looking for a talented Computer Scientist to join our growing development team. Your expertise in data will help us take this to the next level. You will be responsible for identifying opportunities to further improve how we connect recruiters with jobseekers, and designing and implementing solutions. […] Required skills and experience:

- SQL and relational databases;
- Data analysis with R (or Matlab);
- Processing large data sets with MapReduce and Hadoop);
- Real time analytics with Spark, Storm or similar;
- Machine Learning;
- Natural Language Processing (NLP) and text mining;
- Development in C++, Python, Perl;
- Experience with search engines e.g. Lucene/Solr or ElasticSearch advantageous

1. **Data driven approach**
2. **High granularity**
3. **High frequency (updated)**
4. **Focus on important skills**



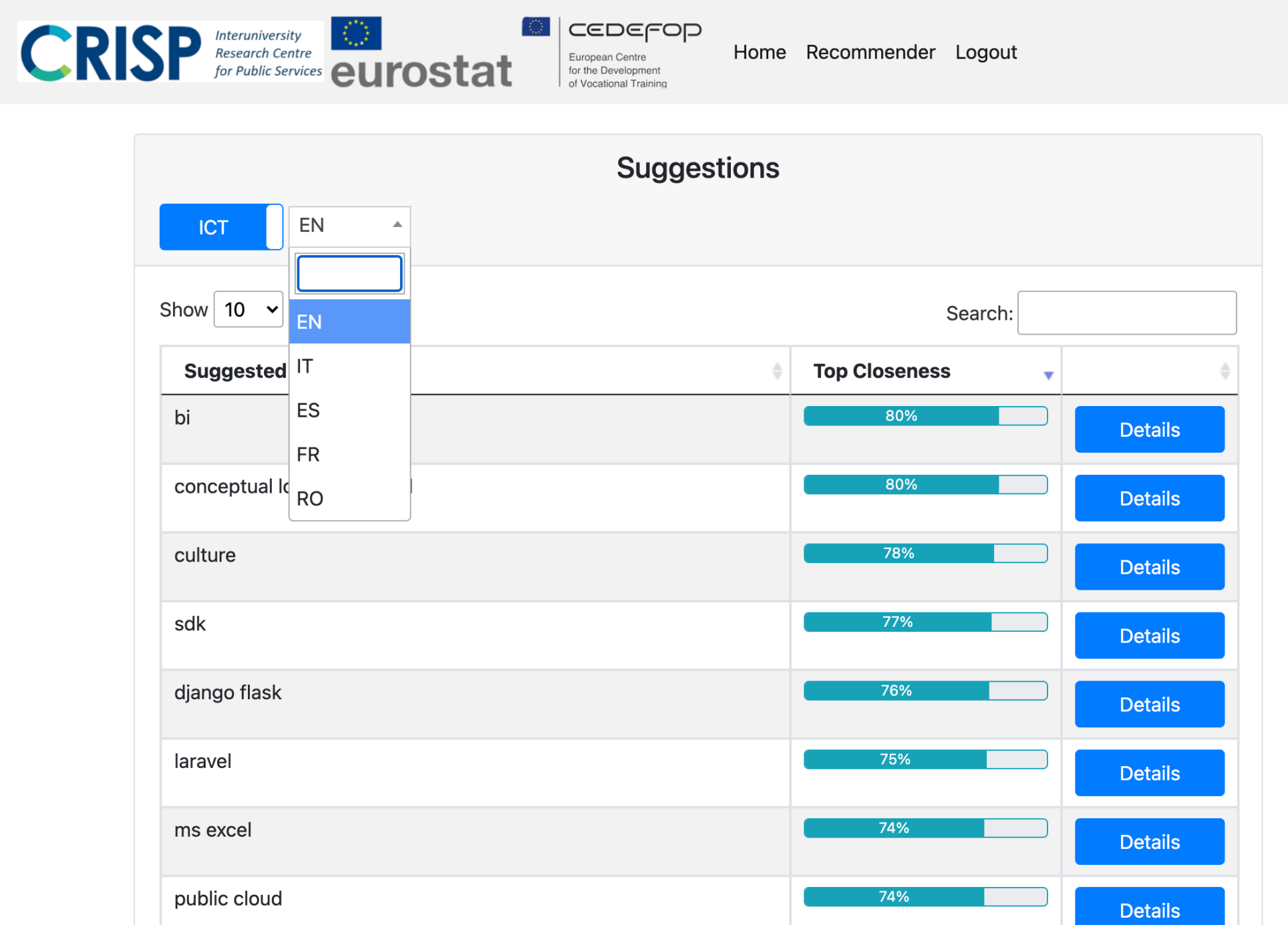
New Emerging Skill: A Human-in-the-loop tool to identify novel skills

We used AI to identify terms that are more likely to represent a novel skills as:

1. They **are used even more** over time

2. Represent something **novel** as it **did not exist before** (in official taxonomies)

3. The system suggests words to a LM **expert** that can **confirm/confute** the machine-suggestion



The LM expert can select the language and the machine suggests words that might represent novel skils

Term: "seo" (ES)

Close to ESCO Skill	Closeness	Model
search engine optimisation	76%	Managers; Professionals
search engines	79%	Managers; Professionals
marketing móvil	52%	Information and communications

How you consider the suggested term?*

A novel skill

What kind of skill is it?*

Hard Skill

How relevant is this suggestion?*

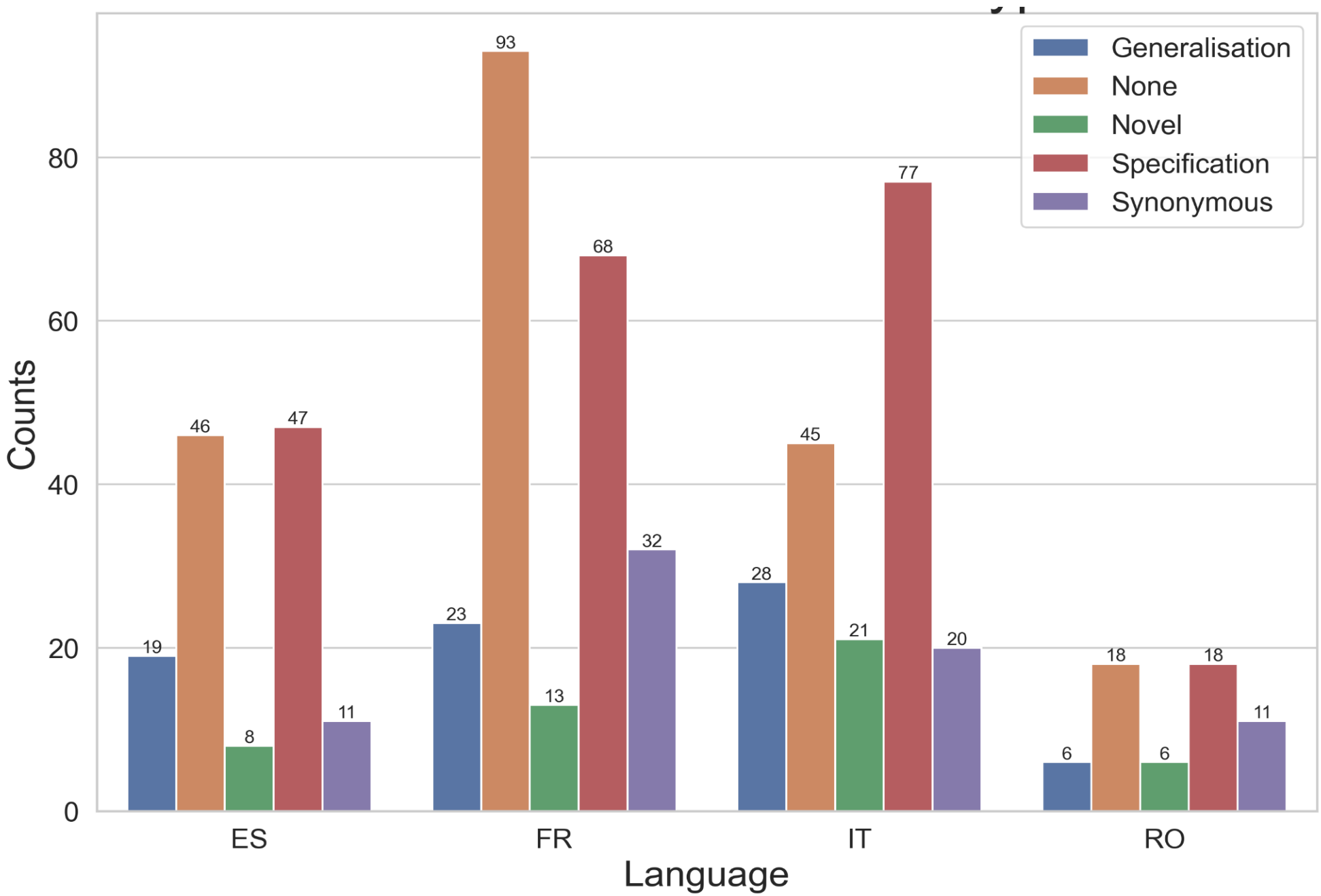
○1 ○2 ○3 ◉4 ○5 ○6

Suggest a correction, if any:

Submit vote

The LM expert is asked to “vote” about the term

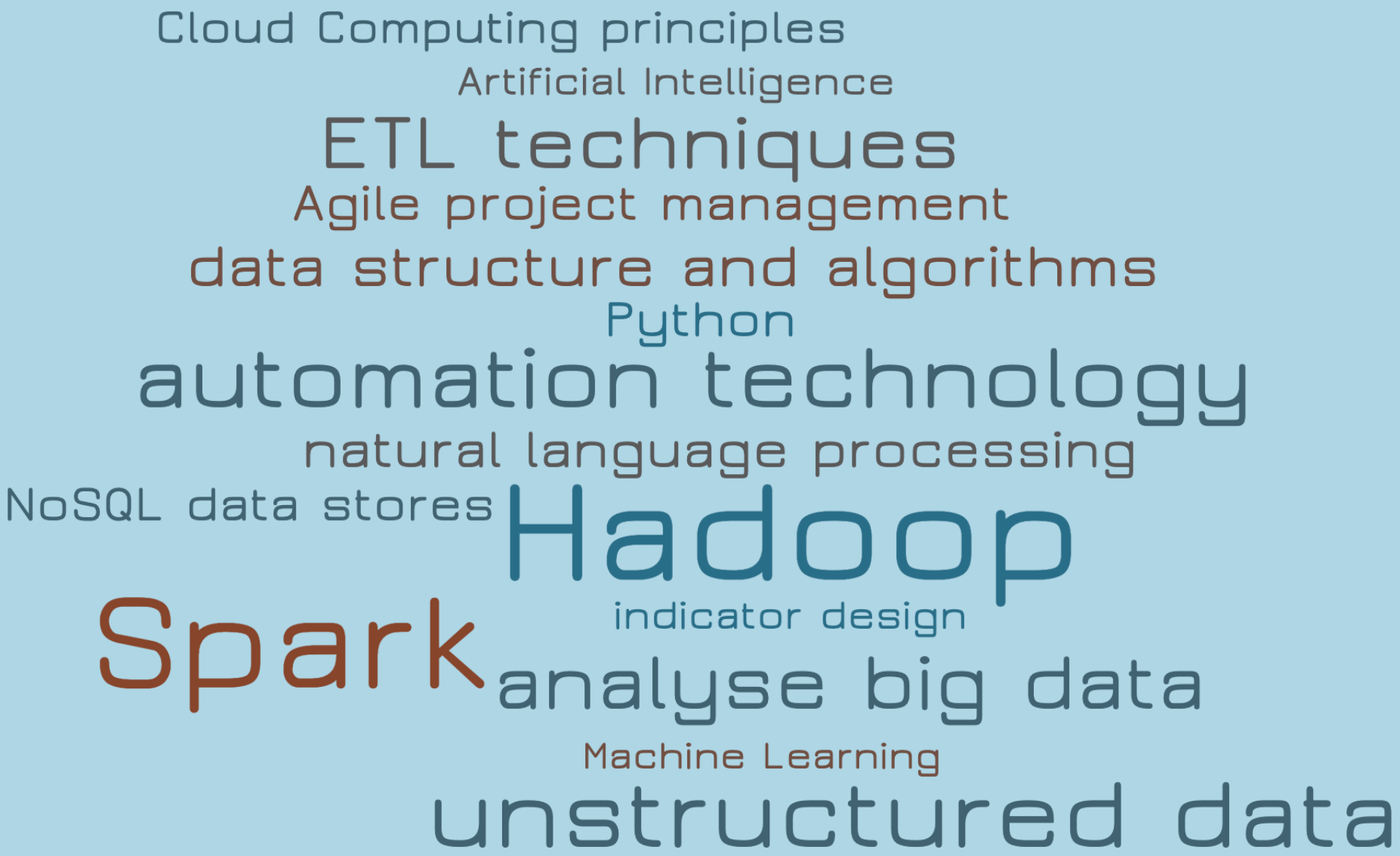
ETE CRISP interuniversity research centre for public services



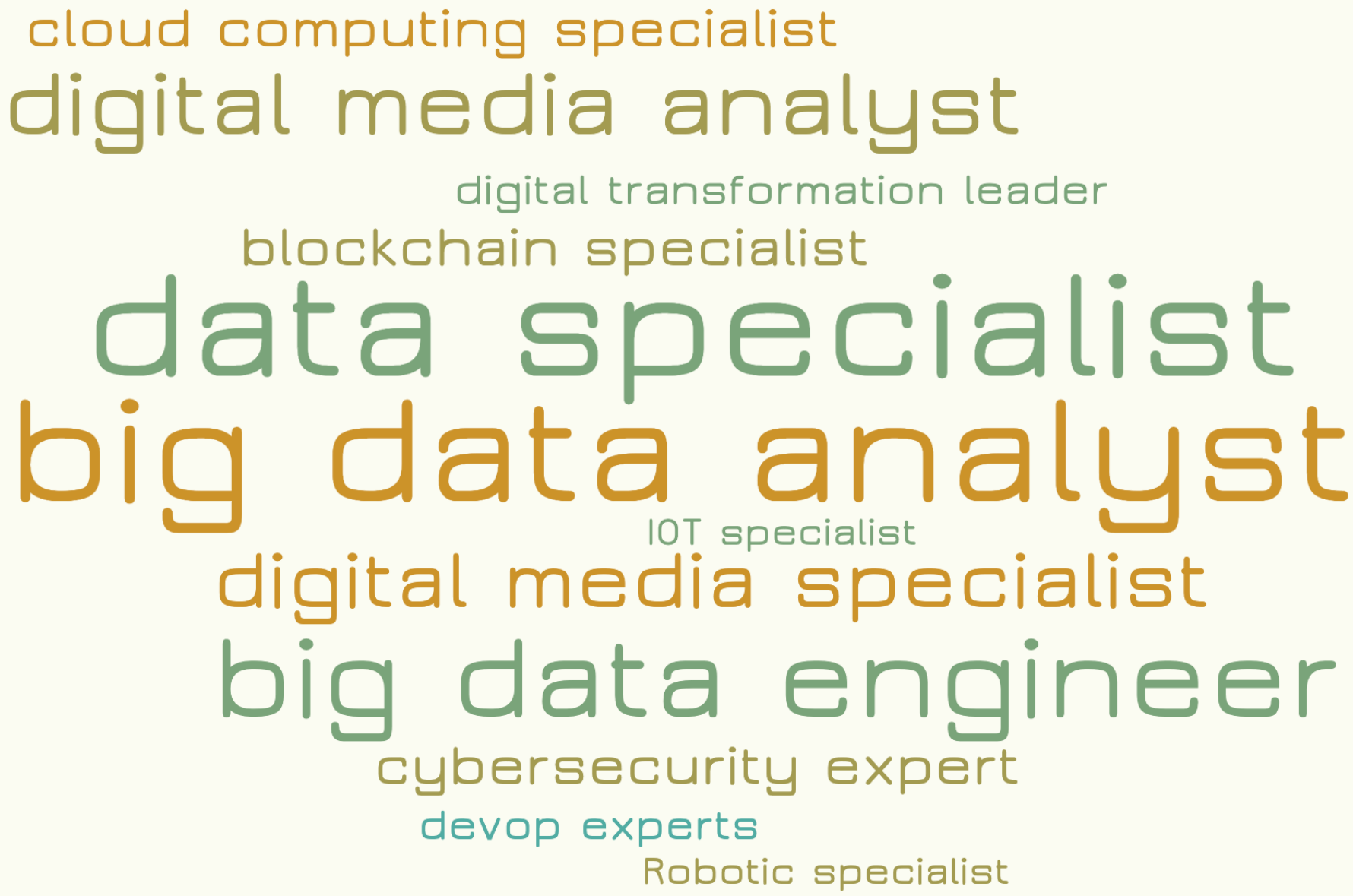
Prototype on ES, FR, IT and RO reveals that the novelty ranges among 6-20% of identified terms and includes synonymous and specifications and generalization of existing ones

CRISP interuniversity research centre for public services

Big data engineer (only novel skills)



Novel occupations



For Citizens and LM Operators

Real-life Application: skill2job

Analysing million Job Ads to measure skill mismatch among demand and offer

Skill2job: Support citizens and operators for upskilling and reskilling

Key elements:

1. *“skills define a job, and not the vice-versa”*
2. *“Asking people to indicate the three skills that they consider as fundamental to do their job, rather than processing an historical CV where the knowledge of a skill might be obsolete or not related with the job at all”*

Tell me your skills: #wordpress #design_graphics #adobephotoshop

Which is your current role? Marketing Specialist

Skill2job: Support citizens and operators for upskilling and reskilling

Tell me your skills: #wordpress #design_graphics #adobephotoshop

Which is your current role? Marketing Specialist

Importance of skills you hold within your job

#wordpress

#design_graphics

#adobephotoshop

Reskilling: you should also acquire:

#webanalytics

#serachengine

Upskilling: You might also be interest

#HTML5

#responsive desig

Career path: similar jobs

Digital Media Specialist

Web and Multimedia Developer

…..

ETF Working together Learning for life

European Training Foundation

For Citizens and LM Operators

Real-life Application: speed of green skills

Analysing million Job Ads to derive a data-driven taxonomy of green skills and analyse the impact of green within jobs

Studying the impact of green economy within jobs and skills

Facts: Environmental impact of jobs is becoming crucial. There is a call to reshape job trying to adapt skills and competences to be sustainable.

Research Questions:

1. Which are the jobs that are more **in demand for green skills**?
2. What characterises a «green job»?
3. Can we estimate the **impact of green skills within a job?** What is the **pervasiveness of green skills** over the known well-established ones?

Countries/Languages: English (EN), German (DE), French (FR), Italian (IT), and Dutch (NL)

ETF Working together Learning for life
European Training Foundation

Skill Pillar - Structure

Occupation Pillar - Structure

Green occupation pillar

ESCO occupation

Occupation at IV level

Green Pervasiveness

Percentage of OJA asking for at least one green skill over the total

Greenness

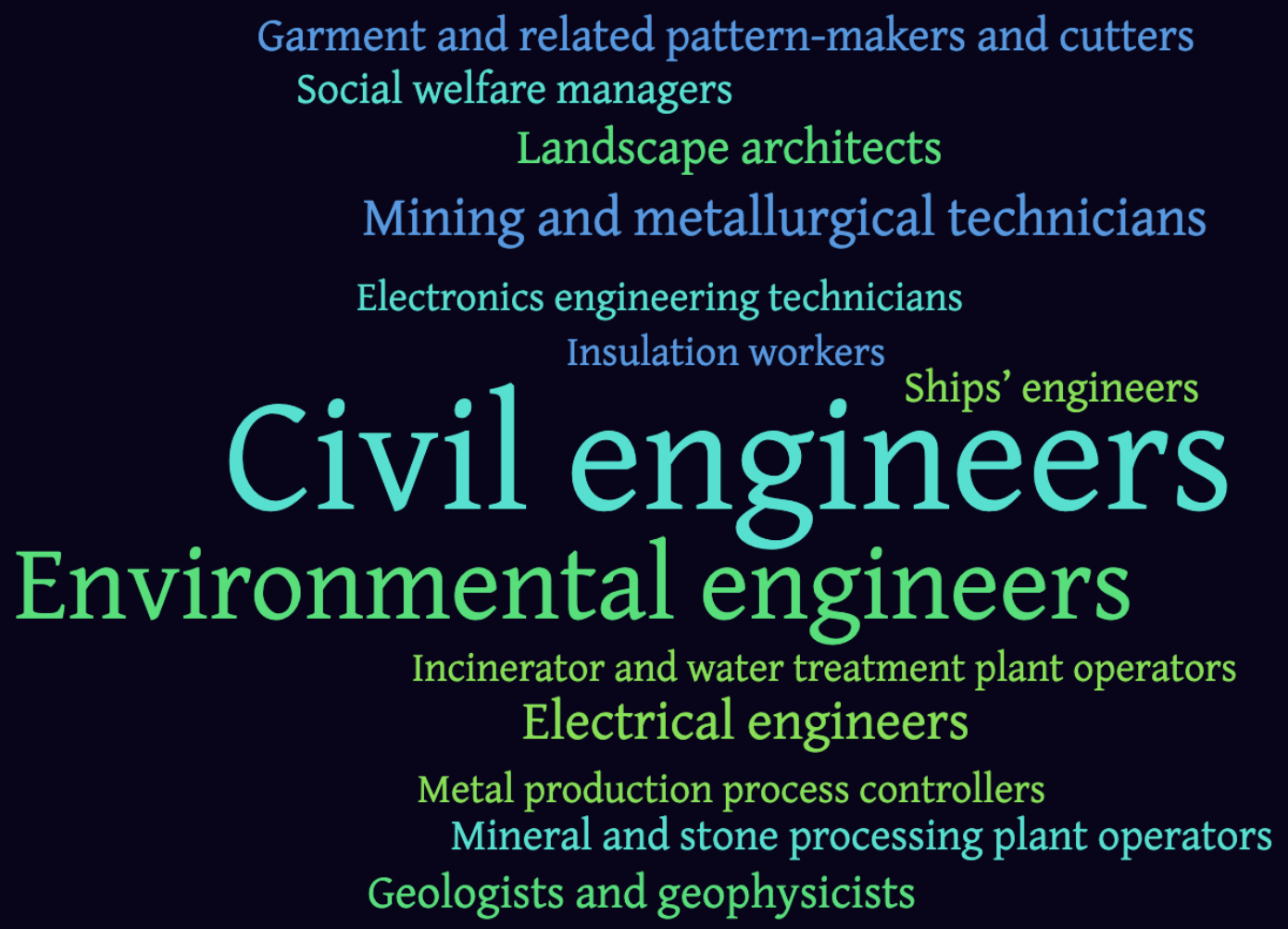
Quota of green skills requested within the occupation with respect to all skills

Green Skill PILLAR

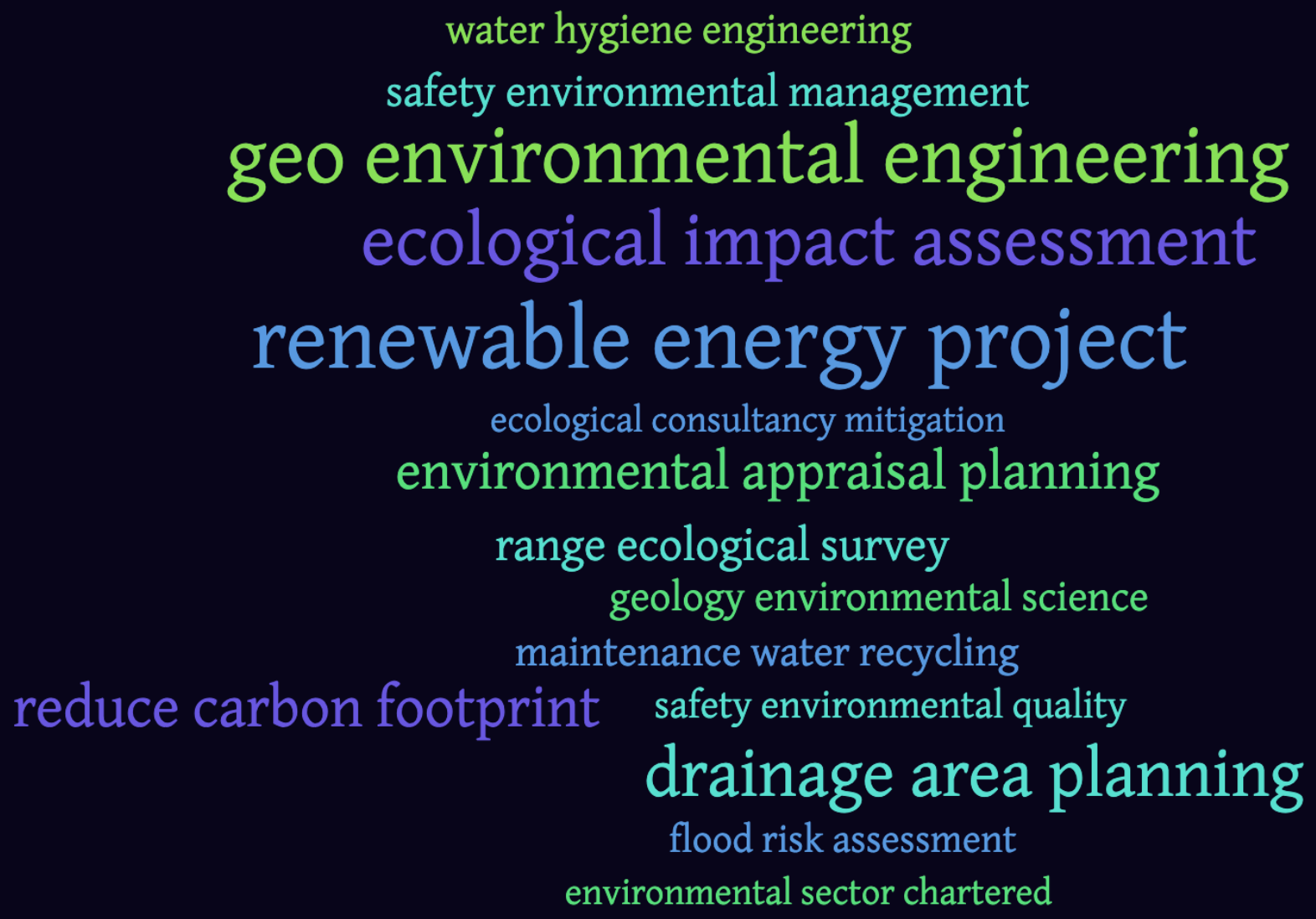
Linkage with top green skills related and relevance

UK-2019 used as a benchmark to compute green pervasiveness and Greenness

Jobs with green pervasiveness > 2%



Top novel green skills identified



Big data does not mean «all data»

Do we really nead to collect data from «any» LM source?

Quality Monitoring:
Source Overlap and Source Replacement Policy

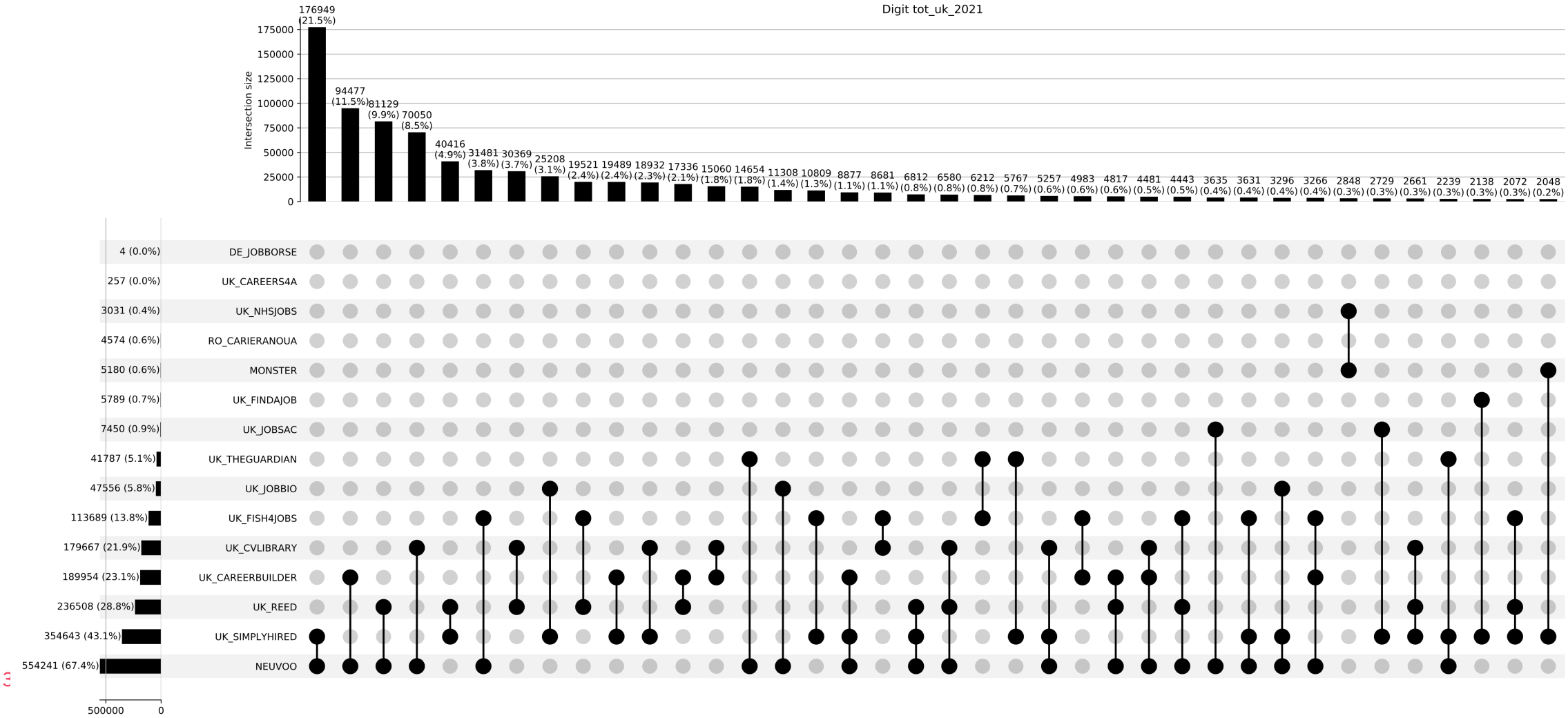
Goal of the activity: to estimate how – and to what extent – sources overalp in terms of duplicated OJAs, and to provide a replacement policy to deal with bizantine sources

Context: A selection of countries to be analysed

Output:

1. Analysis of overlap country-by-country
2. Compute the marginal contribution of sources (i.e., the minimal list of sources to be included that maximises the coverage)

Quality Monitoring:
Source Overlap and Source Replacement Policy



Quality Monitoring:
Source Overlap and Source Replacement Policy

Num Sources	Coverage	Marginal Contribution
2	60,4	39,6
3	69,9	30,1
4	77	23
5	82,2	17,8
6	87,4	12,6
7	91,4	8,6
8	94,7	5,3
9	96,1	3,9
10	96,8	3,2
11	97,9	2,1
12	99,1	0,9
13	98,4	1,6
14	98,9	1,1
15	100	0

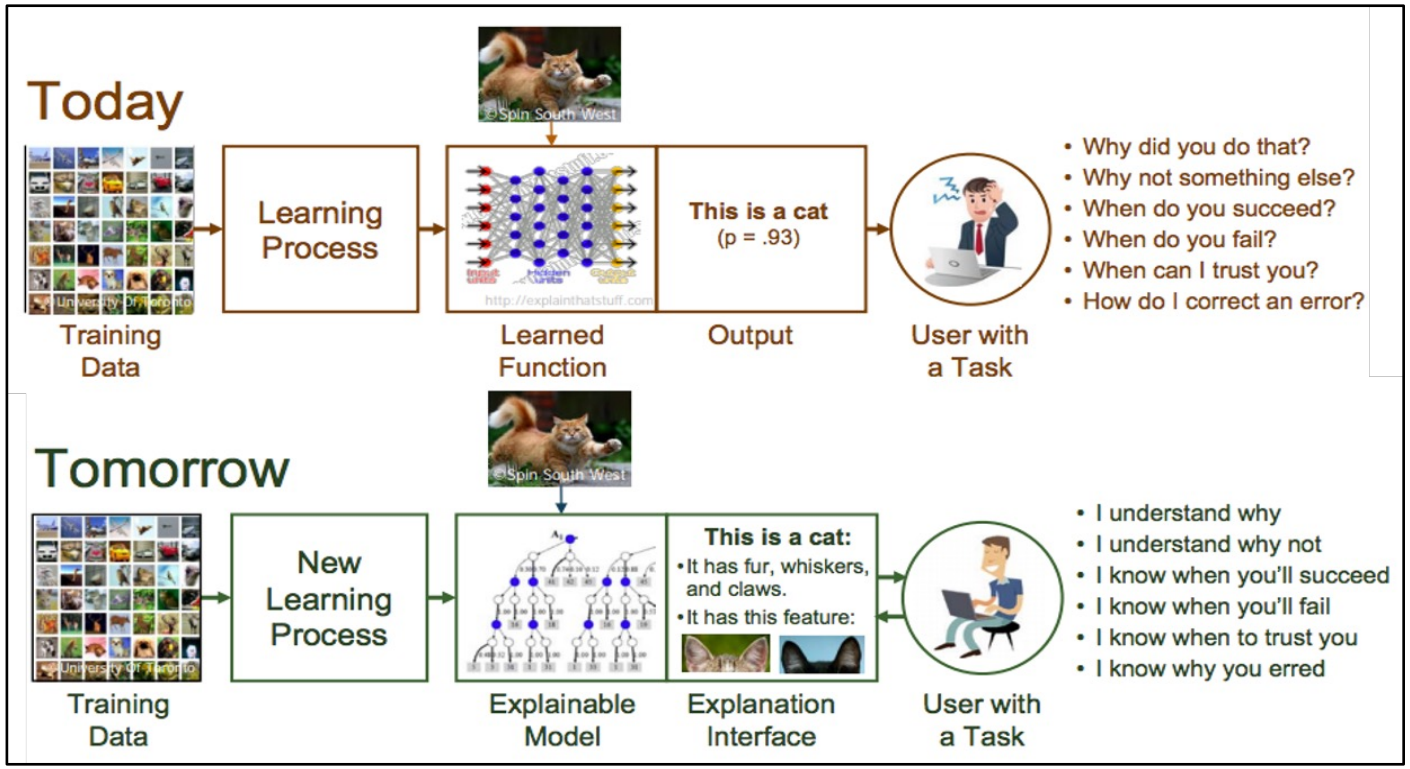
Look! For UK-19 7 sources out of 15 are enough to cover 90+% of Web labour market

Can we trust ML results?

Is there a way to “unbox” the box, to better understand the rationale behind the scenes?

eXplainable AI (XAI)

The notion of AI models as black boxes strongly limits their acceptance in several fields.



2017, American Defence Advanced Research Projects Agency (DARPA): toolkit of machine learning techniques to produce more explainable models to garner human users' trust and foster adoption of AI

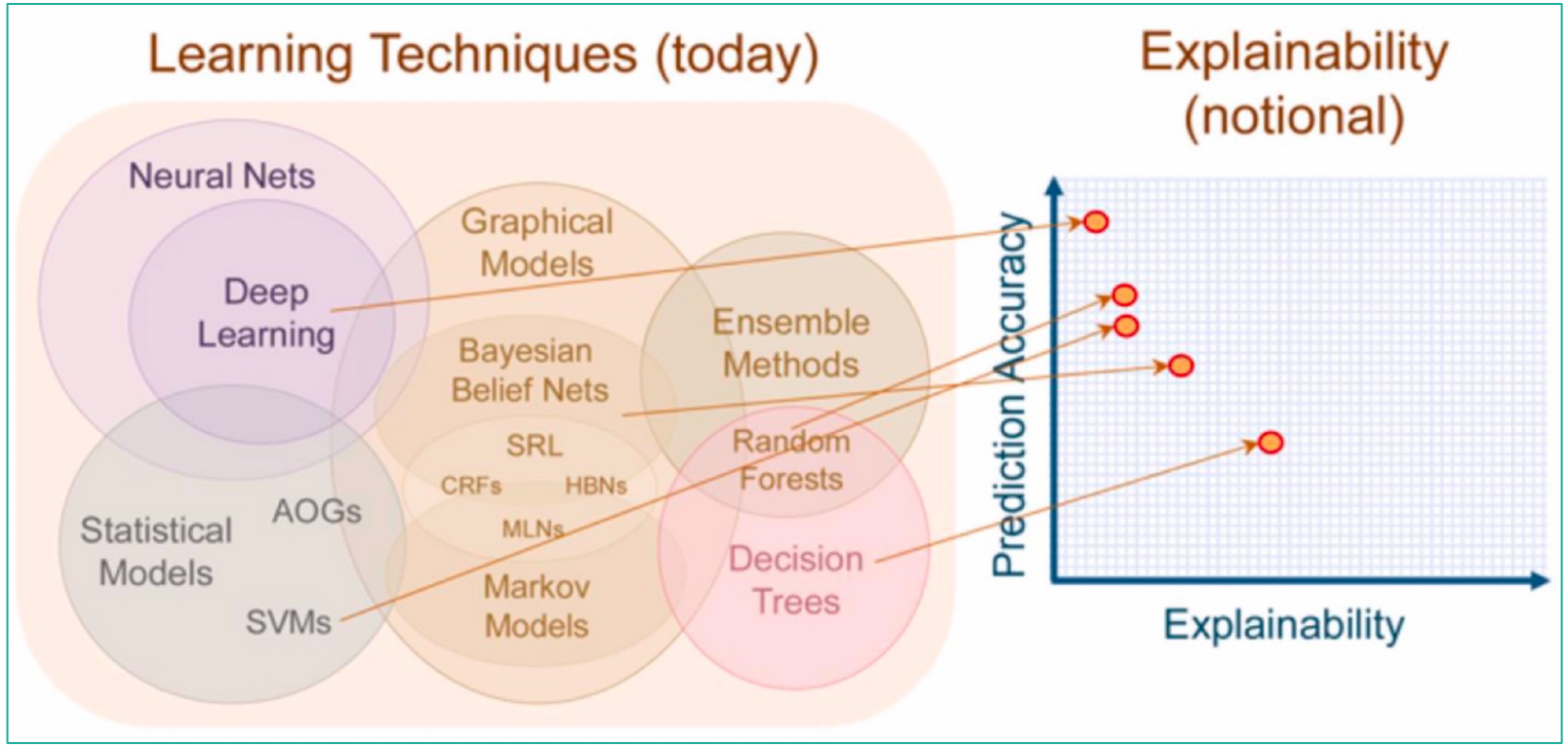
2018, GDPR (General Data Protection Regulation), defines the right to explanations (Art. 13-15), asking the data controllers to provide data subjects with information about existing automated decision-making processes

ETF Working together Learning for life
European Training Foundation

eXplainable AI (XAI)

Complex models (such as deep neural networks or support vector machines): increased prediction performances but is not explainable.

There is a trade-off between performances and explanability.



Once a ML-based classifier is updated over time…

Questions

1. Can we estimate to what extent $M2$ classifies coherently to the past predictions made by $M1$?
2. Why does the criteria used by $M1$ now leads M2 to classify on a different class?
3. Can natural language summarise those differences to final-users that do not have technical competences?

Once a ML-based classifier is updated over time…

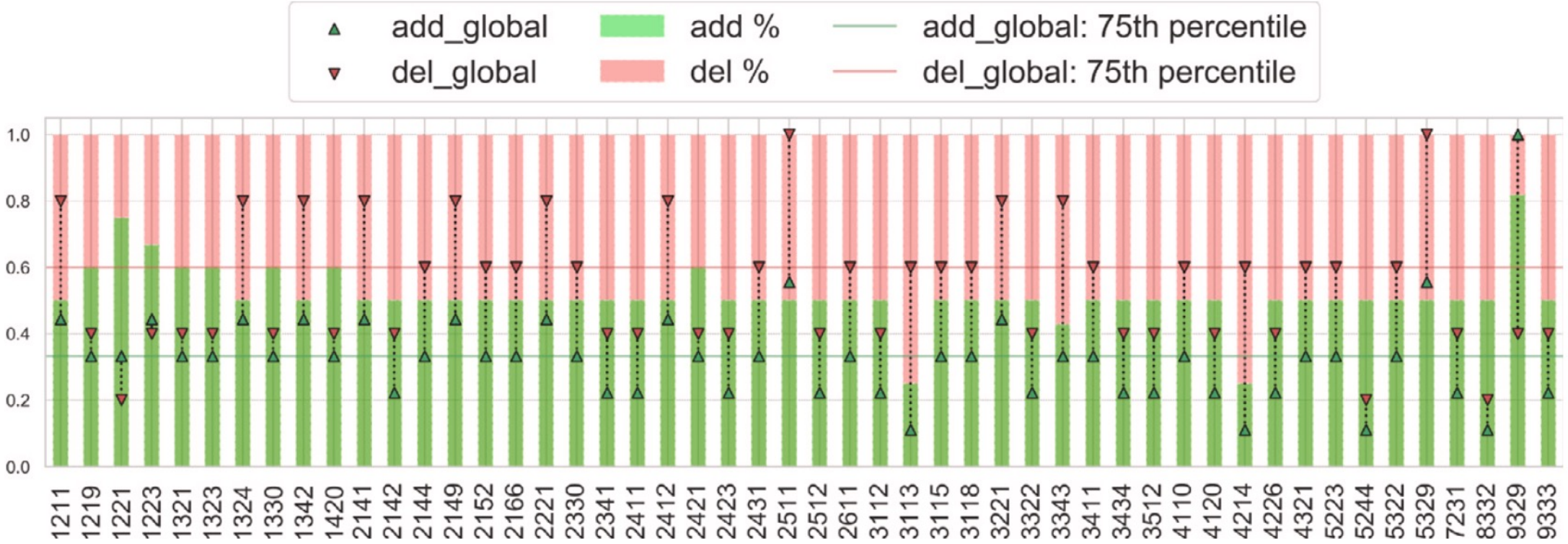
```
The model now uses the following classification rules for this class:
This class has 6 added classification rules, but only 3 are used to classify the 80%
of the items.

 - Having Engineer.
 - Having Developer but not Engineer.
 - Having JavaDeveloper but not Engineer, and Developer.

The model is not using the following classification rules anymore:
This class has 2 deleted classification rules.

 - Having WebDeveloper but not Developer, Analyst, Systems, JavaDeveloper, and NetDeveloper.
 - Having Software but not Developer, Analyst, Systems, WebDeveloper, JavaDeveloper, and NetDeveloper.

There are no 'unchanged' classification rules.
```

Fig. 10. The ContrXT output for *2512, Software Developers* vs *2511, Systems Analysts* in the class vs class case using a DT to explain the RF model of Table 2.

ContrXT is a tool developed by CRISP, see contrxt.ai

ETF Working together Learning for life
European Training Foundation

Once a ML-based classifier is updated over time…



ContrXT is a tool developed by CRISP, see contrxt.ai

ETF Working together Learning for life
European Training Foundation