Big Data for Labour Market Intelligence

Capacity development programme 2022

Module 1: Technical training

Session 9

Use of Big Data for different analyses of occupational and skills dynamics in the labour market.
Combination of data from different sources

Speaker: Mauro Pelucchi

29/11/2022

European Training Foundation

Data production system

Tunisia (04/2020 to 07/2022)

680,191 OJVs - > 175,203 deduplicated

Ukraine general (04/2020 to 07/2022)

2,571,655 OJVs - > 1,304,262 deduplicated

Georgia (04/2021-07/2022)

129,271 OJVs - 84,817 deduplicated

Egypt (new)

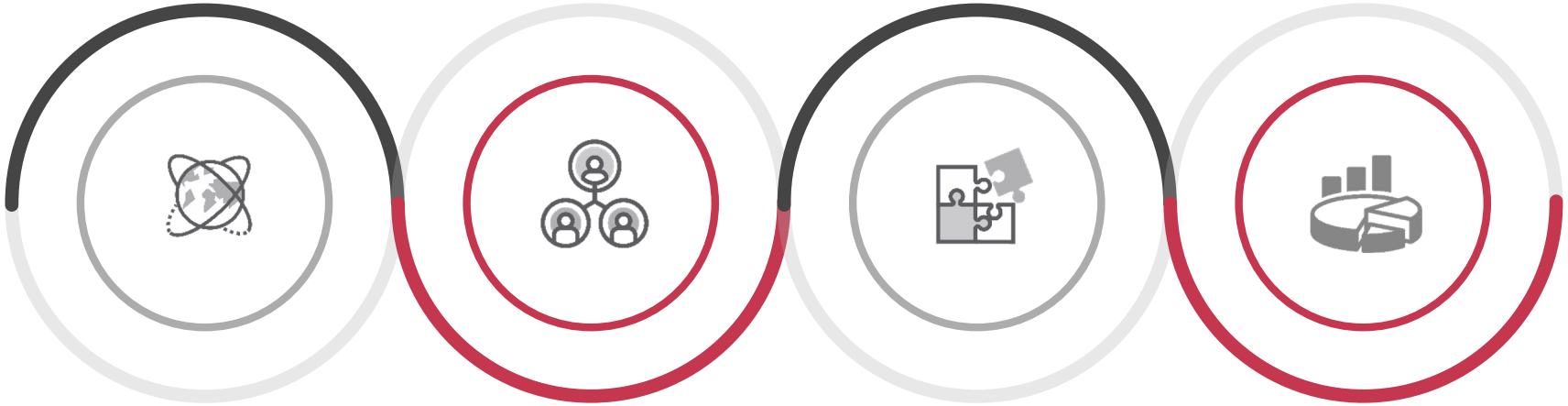
1,307,678 OJVs – 391,701 deduplicated

Kenya (new)

(collection started in september 2022)

Collecting and decoding labor market data

Real-time job market data offer up-to-date insights not possible through traditional sources



Capture job market data

Tagging and structuring

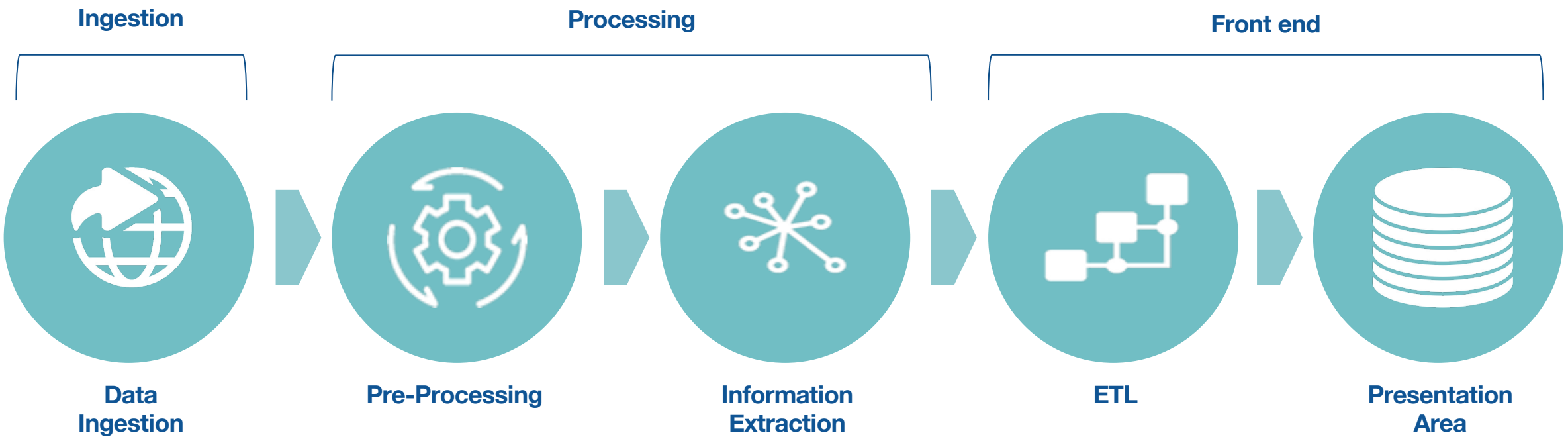
A common language

Data ontology allows for comparisons

Drawing conclusions

Insight from in-demand skills and real-life career patterns

Overall Data Flow



Data Ingestion

- **Goal:**
  - Collect massive amount of heterogeneous data from several unpredictable number of sources
- **Challenges:**
  - Guarantee data completeness and consistency
- **Approach:**
  - Develop a multi-technique framework (crawling, scraping, API) to fit different website characteristics: volume of vacancies, different languages, technology structure, non-invasive approach and policy agreements
  - Prevent data sources losses, via redundancy policies
  - Detect and collect metadata to improve information value
- **Features:**
  - To grant process governance in data Ingestion phase, we developed a monitoring system with a scheduling tool and an alerting module

Data Pre-Processing

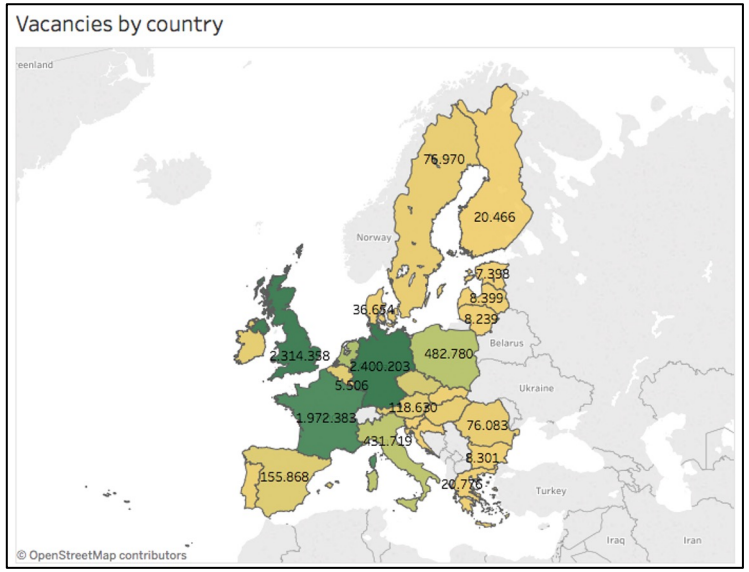
- **Goal**
  - **Feed information extraction phase with proper data**
- **Challenges**
  - **Measure, monitor and increase Data Quality, to maximize completeness, consistency, complexity, timeliness and periodicity**
- **Approach**
  - **Develop a multi-phase pipeline, focused on:**
    - **Vacancy Detection: analyze website page to select only content referred to vacancies**
    - **Deduplication: detect duplicated vacancy posts to obtain a single vacancy entity**
    - **Date detection: identify release and expire dates through vacancy description analysis**
    - **Vacancy duration: method to define expire date, when not explicitly available**
- **Features**
  - **Guarantee Data Quality during all processing phases**

Data Classification

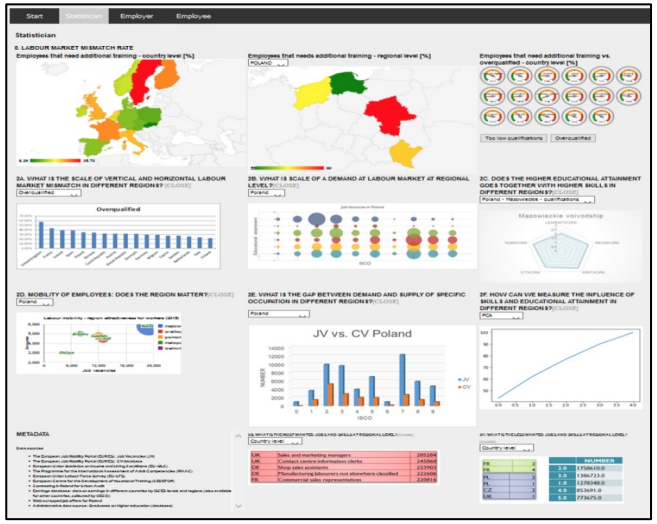
- **Goal:**
  - **Extract and structure information from data, to be provided to the presentation layer**
- **Challenges:**
  - **Handle massive amount of heterogeneous data written in different languages**
- **Approach:**
  - **Develop an adaptable framework, language dependent, tailored on different information features. Some relevant challenges:**
    - **Occupation feature classification: combined methods such as Machine Learning, Topic Modeling and Unsupervised Learning**
    - **Skill feature classification: another different combined methods, such as Text Analysis with corpus based or Knowledge based similarity**
- **Features:**
  - **Guarantee Explainable information extraction, logging classification methods and relevant features.**

Presentation

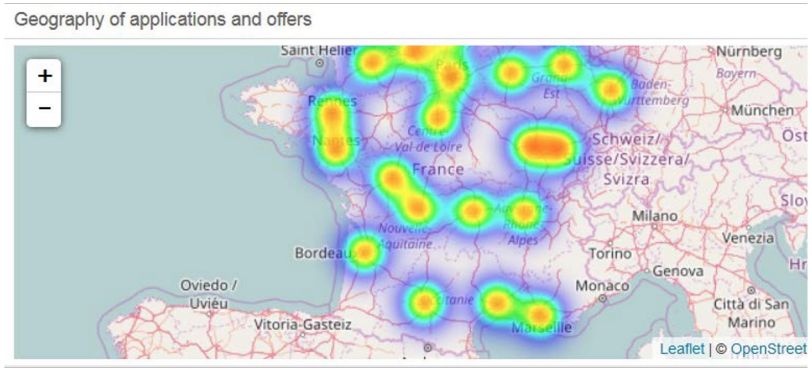
- **Goal**:
  - Provide data in an Analytic environment, to support users across their different analysis needs
- **Challenges:**
  - Personalized data access to different types of stakeholders
- **Approach**:
  - Develop different navigation paths, for different users:
    - **Data scientists/analysts**: access data via a Discovery Lab environment, to ensure free data analysis needs
    - **Decision makers**: access information through storytelling and dynamic dashboards



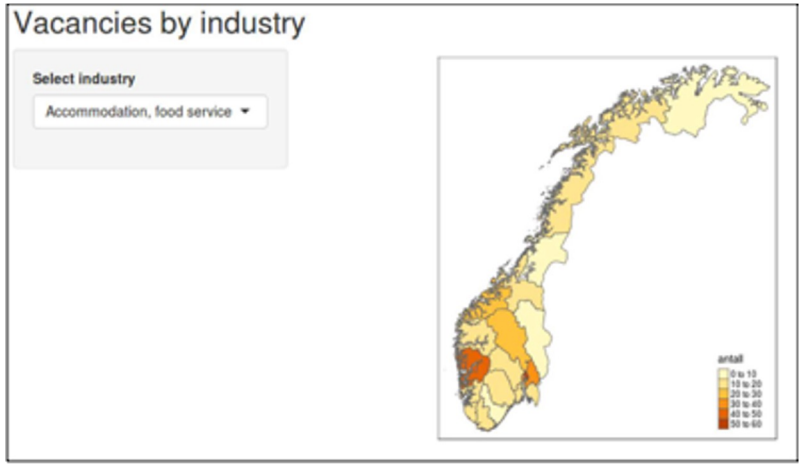
Dashboard with point and click capabilities



Labour Market Portal
(from 2017 Big Data Hackathon – Poland Team)



Map (from 2017 Big Data Hackathon – French Team)



Dashboard (from 2017 Big Data Hackathon – Norwegian Team)

Lab session
Revealed comparative advantage

Lightcast

RCA

Given a set of occupations $\bar{O} = \{o_k, k = 1, \ldots, m\}$, a set of skills $\bar{S} = \{s_j, j = 1, \ldots, p\}$, and a matrix $\underset{m \times p}{M}$ that contains a value of sf for each pair of occupations $o_k \in \bar{O}$ and skills $s_j \in \bar{S}$, rca for o_i and s_l is defined as:

$$rca(o_i, s_l) = \frac{sf(o_i, s_l) / \sum_{j=1}^{p} sf(o_i, s_j)}{\sum_{k=1}^{m} sf(o_k, s_l) / \sum_{k=1}^{m} \sum_{j=1}^{p} sf(o_k, s_j)}$$

Occupation similarity

From the notebook:

- We will select a country
- For the country we will calculate the RCA
- And we will see our create bridge between occupations

Cmd 1

RCA

In this notebook we will see the computetion of RCA and now use it to calculate a KPI about the distance from occupations.

Thanks to Simone Perego for the Notebook

Cmd 2

```
%md
select idcountry, idesco_level_4, esco_level_4, idescoskill_level_3, escoskill_level_3, count(distinct general_id) as num_ojv from ft_skill_analysis_en  where idesco_level_4 in ('1330', '2511', '2512', '2513', '2514', '2519', '2521', '2522', '2523', '2529', '3511', '3512', '3513', '3514', '3521', '3522') and year_grab_date = 2020 and idcountry in ('IT', 'DE', 'FR', 'ES', 'IE', 'CZ') and idescoskill_level_1 <> 'No digital'
group by idcountry, idesco_level_4, esco_level_4, idescoskill_level_3, escoskill_level_3
```

Cmd 3

```
%sql
select * from default.data_rca_csv limit 10
```

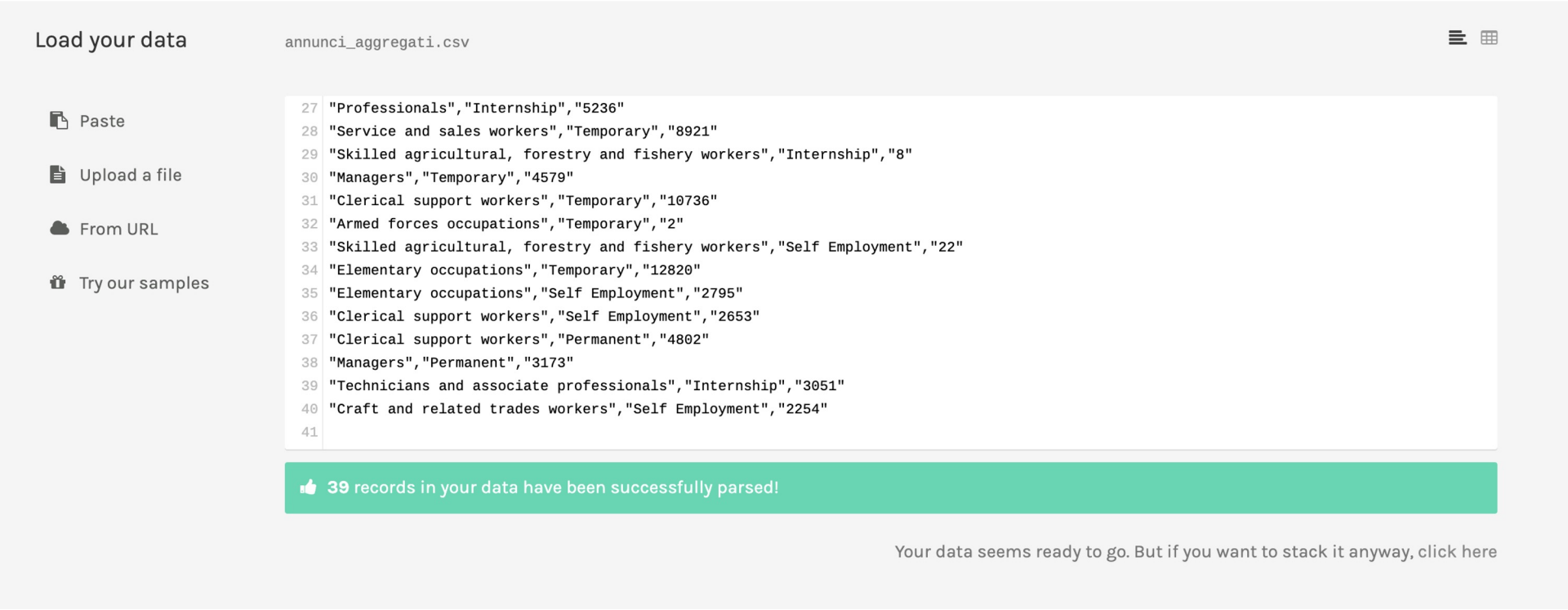
(1) Spark Jobs

	idcountry	idesco_level_4	esco_level_4	idescoskill_level_3	escoskill_level_3	num_ojv
1	FR	1330	Information and communications technology service managers	STACK_157	elasticsearch	112
2	CZ	2511	Systems analysts	ESCOv1_10465	use digital device operating systems	116
3	ES	3512	Information and communications technology user support technicians	ESCOv1_3737	manage system security	272
4	DE	3513	Computer network and systems technicians	ESCOv1_529	cloud technologies	6469
5	FR	1330	Information and communications technology service managers	STACK_40	eclipse	65
6	DE	2523	Computer network professionals	ESCOv1_4128	quality standards	1980

Showing all 10 rows.

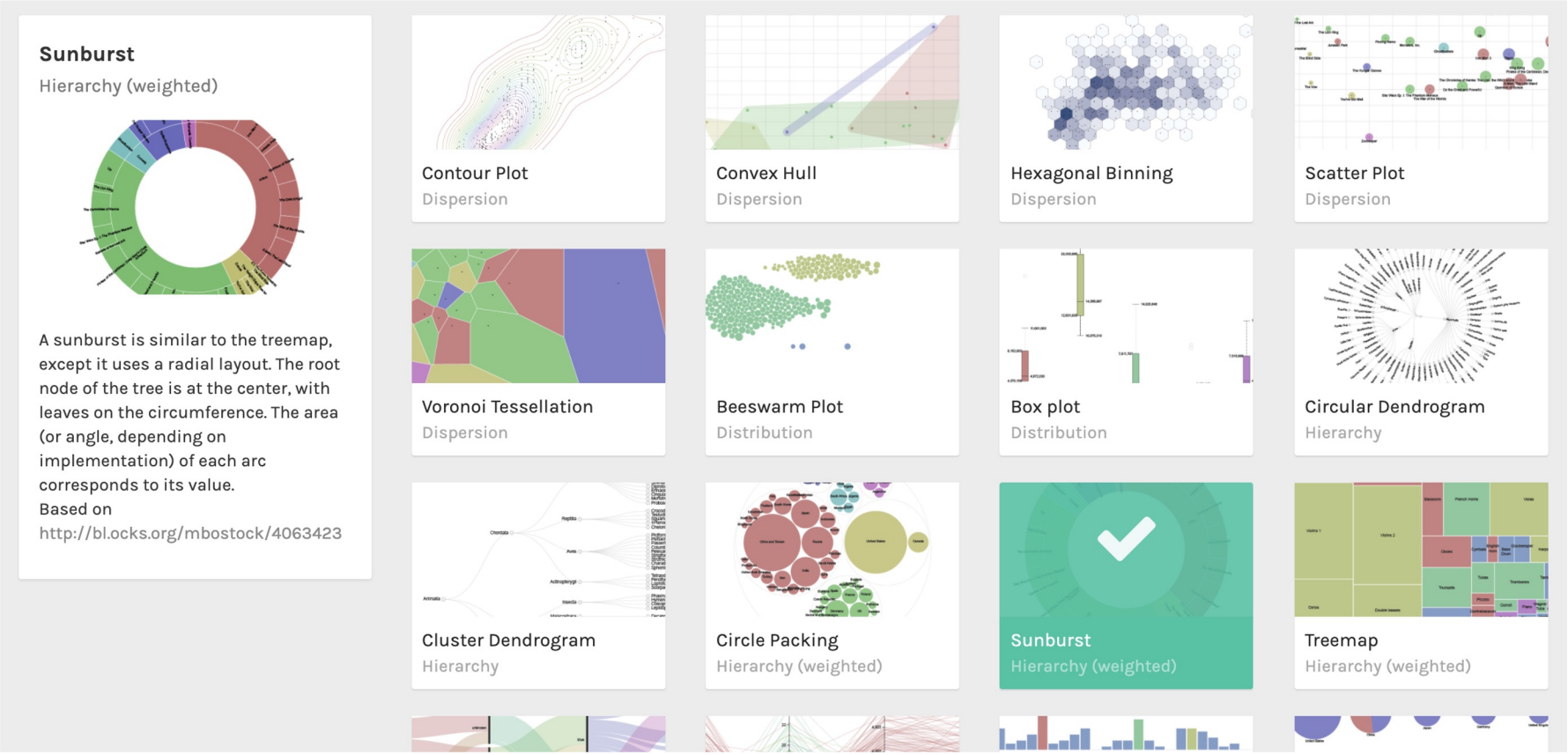
rawdata.io

https://app.rawgraphs.io/



rawdata.io

https://app.rawgraphs.io/

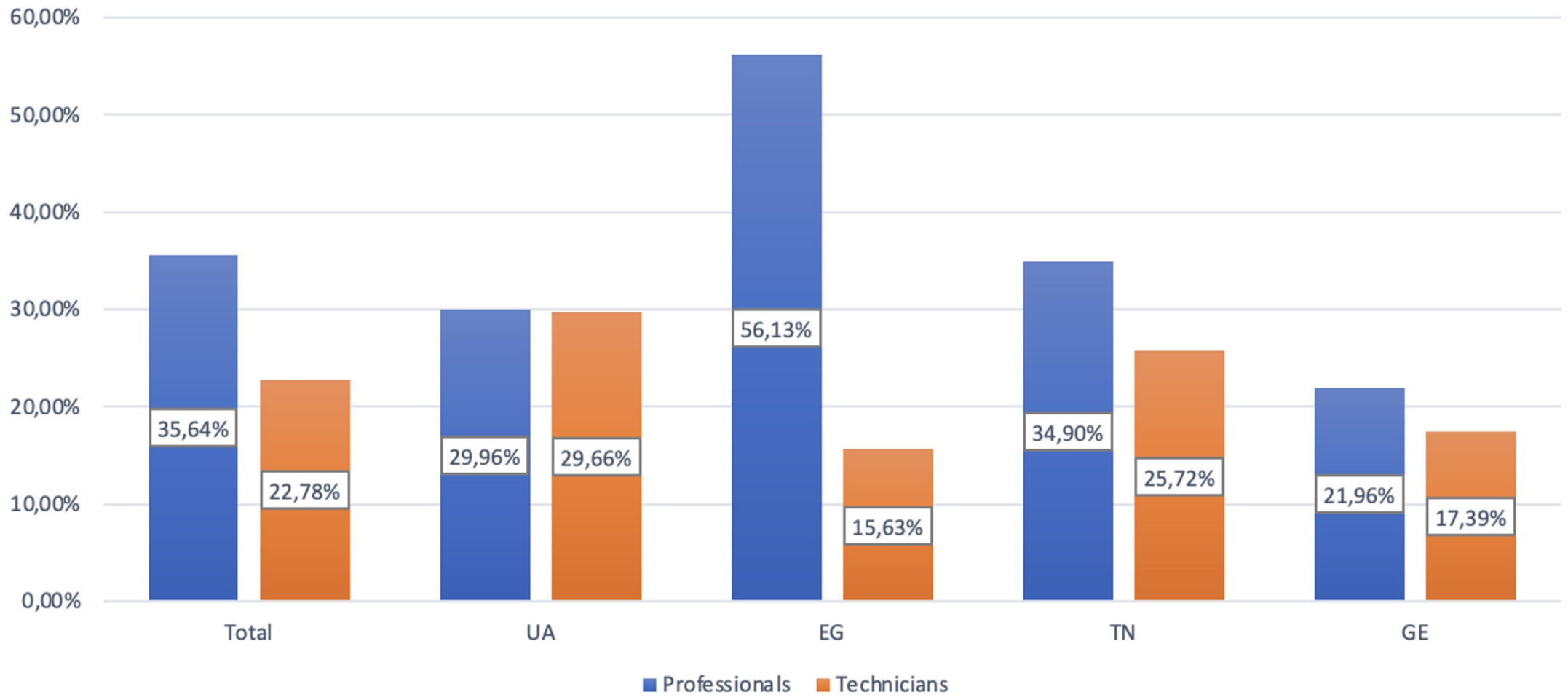


ETF Working together Learning for life
European Training Foundation

Skills and occupations

Lightcast

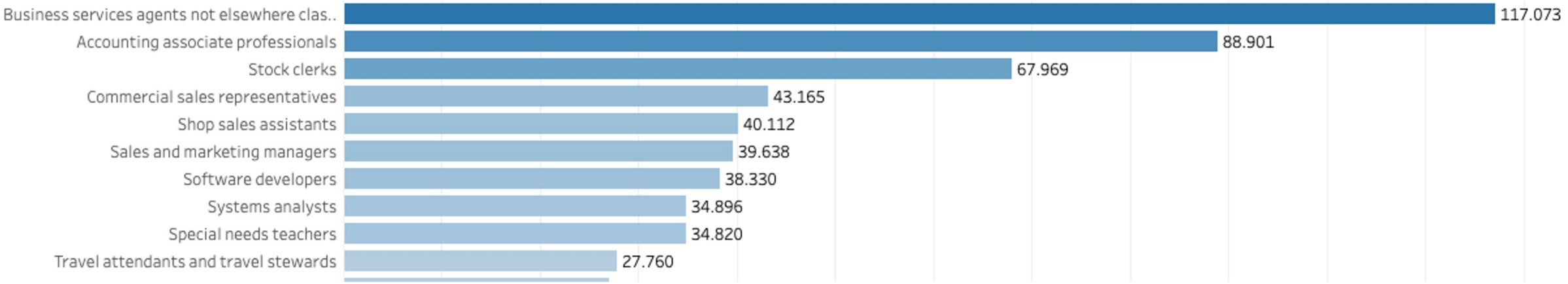
Strong role for professionals



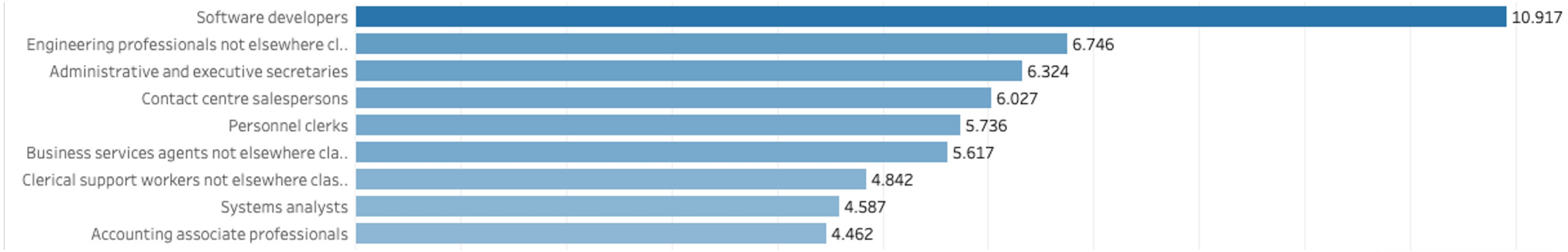
Source: ETF Dataset - 2020-2022 Sept

Detailed occupations

UA - ETF Dataset - Unique Job postings 2020-2022



TN - ETF Dataset - Unique Job postings 2020-2022



Skills and ESCO

ESCO taxonomy allows to group skills into categories. Major distinction between hard and soft skills.

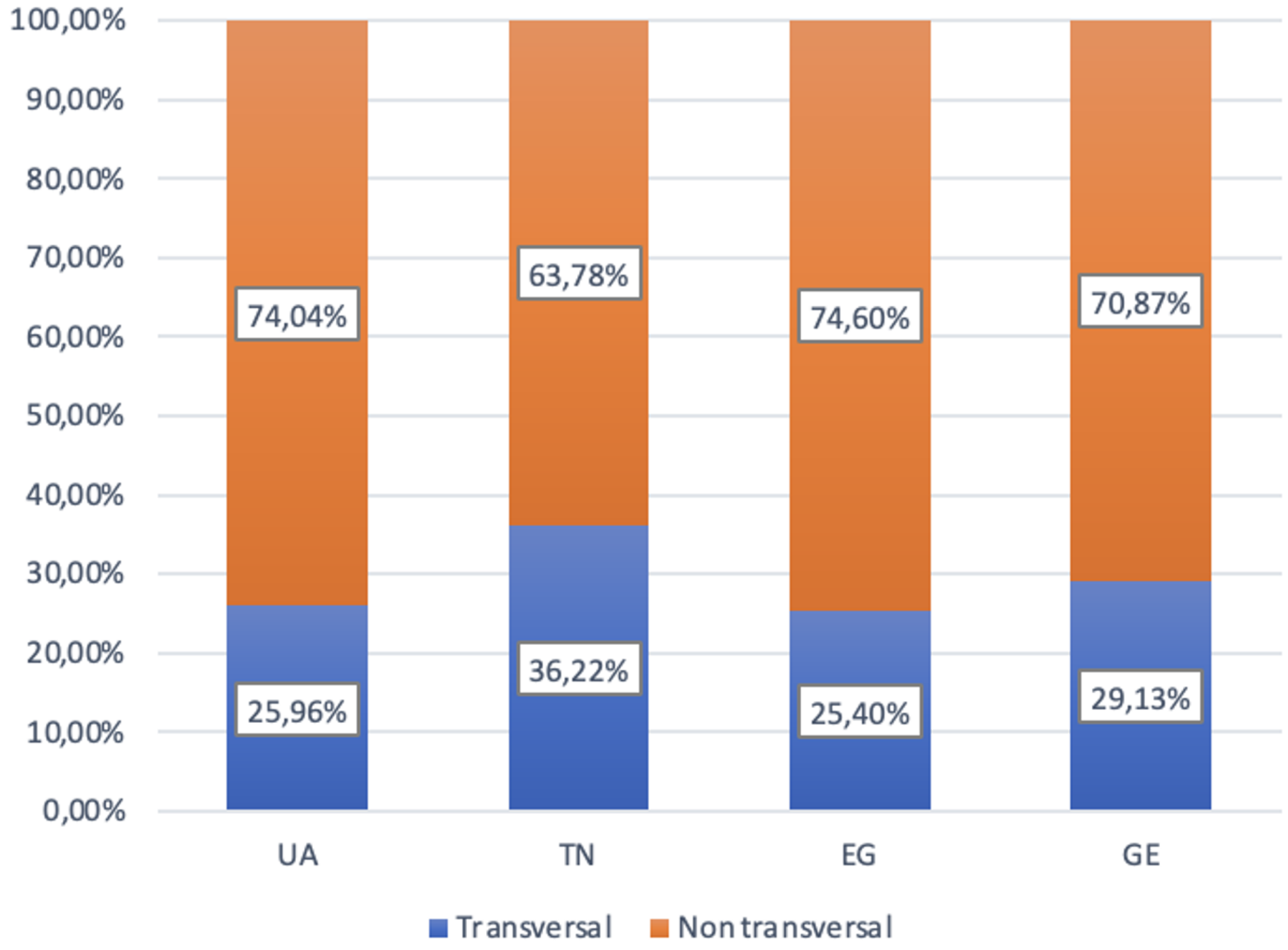
- **Hard skills:** job-specific skills and competences that are needed to perform a specific job or task (examples are knowledge of specific software or instruments, specific manual abilities etc.)
- **Soft skills:** transversal in nature and refer to the capacity of individuals to interact with others and the environment (examples are communication skills, problem solving etc.).

Transversal skills in ESCO

- **Thinking skills.** Ability to apply mental processes and reasoning to solve complex problems, to increase knowledge and to perform complex tasks.
- **Social interaction.** Ability to develop interact and engage with colleagues, clients and customers.
- **Application of knowledge.** General application of skills commonly used in the workplace and in learning; knowledge of the organization and the working environment.
- **Attitudes and values.** Person's work style, preferences and work-related beliefs that underpin behaviour. Examples are ability to adapt to change, to work independently, to meet commitments etc.

We add to transversal skills also Languages

Skills: transversals vs non



ETF Working together Learning for life
European Training Foundation

Most relevant transversal skills - UA

Most relevant non transversal skills - UA

Labour Market Challenging factors

Lightcast

Labour Market – challenging factors

- Skills Evolution
- New Emerging Occupations
- Job Automatisation/Replacement
- …..

Compute Skills Rates and Relevance

Skills Rate estimate the incidence of Digital, Hard non Digital and Soft skills in a single occupation;

Skills Relevance defines the weight of each skill for the considered occupation

Idea: Exploit the informative powerof the OJV for computing the Skill Rate and skill Relevance

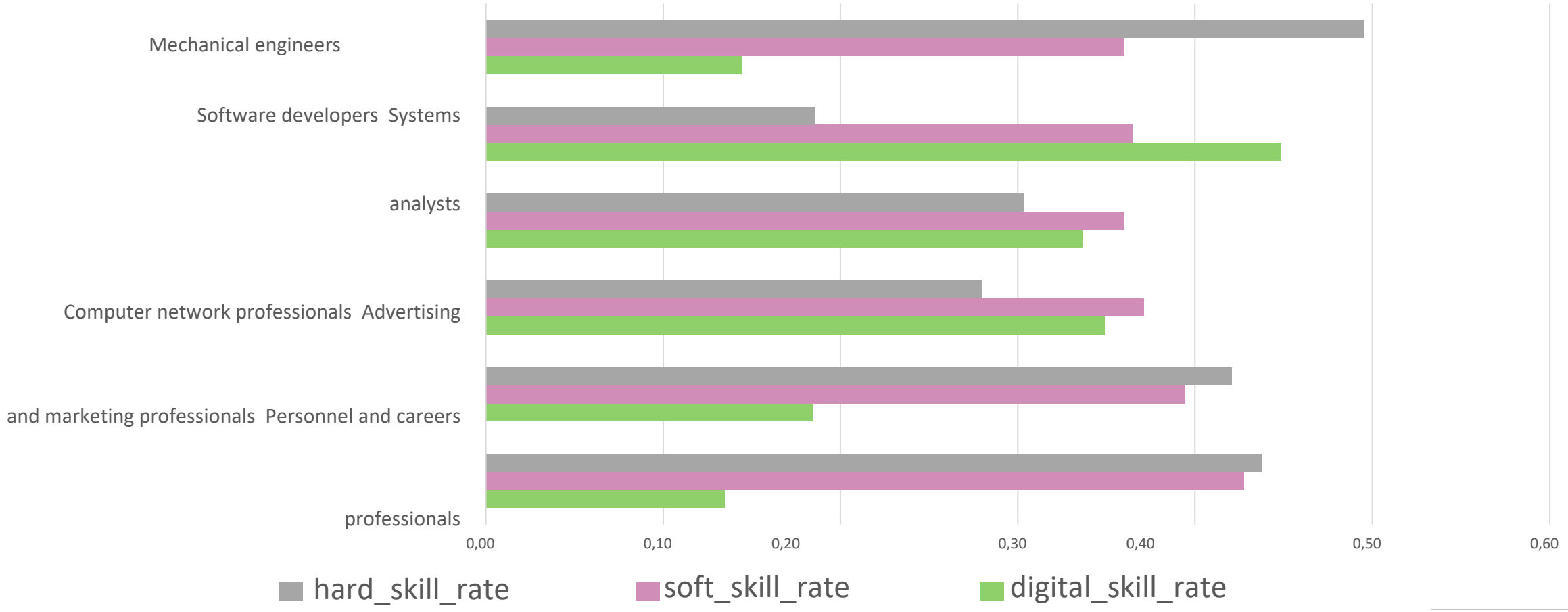
Skills Rate and Relevance methodology was defined and used by CRISP in 2017 and 2018 for the Italian Observatory of Digital Skills promoted by the Italian ICT associations

Definition and computation method

$$skills\ \text{Rate} = \frac{frequency\ of\ skills\ (digital\ or\ hard\ or\ soft)}{frequency\ of\ digital + hard + soft\ skills}$$

$$skills\ relevance = \frac{n°\ of\ vacancies\ of\ (occupation, skill)}{n°\ of\ vacancies\ of\ (skill)} \times \frac{n°\ of\ vacancies\ of\ (occupation, skill)}{n°\ of\ vacancies\ of\ (occupation)}$$

Skills Rate for some occupations



Final remarks

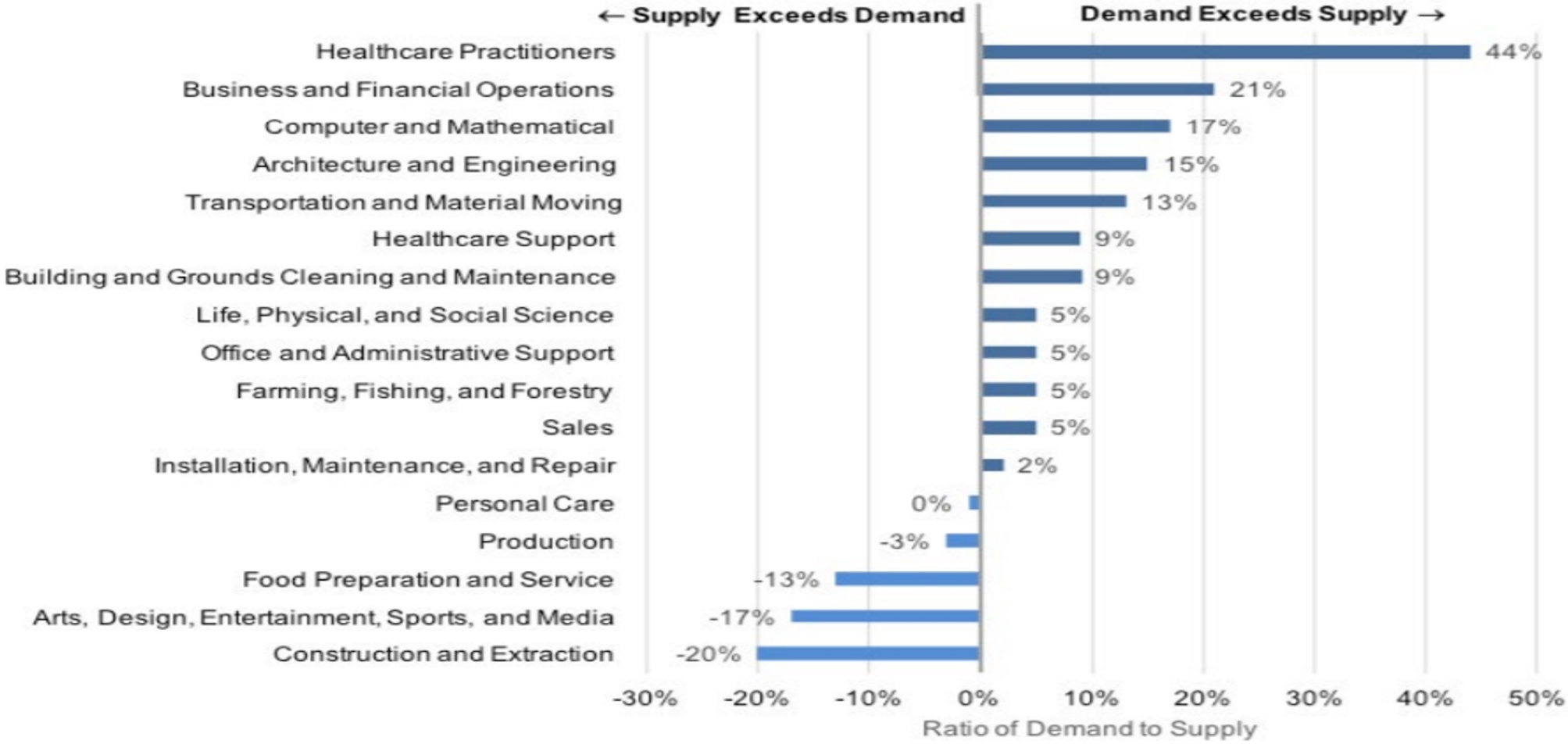
- Knowledge become crucial to observe the challenging factors of LM:
  - Skills Evolution
  - New Emerging Occupations
  - Job Automatisation/Replacement
  - …
- OJV are complementary to the other sources that we normally use to understand LM phenomena
- We are only at the beginning of exploiting the informative potential of OJV - Big Data

Big Data Insights

Lightcast

Not All Shortages are Gaps

Deeper Characterization of the Skills Gap



European Training Foundation

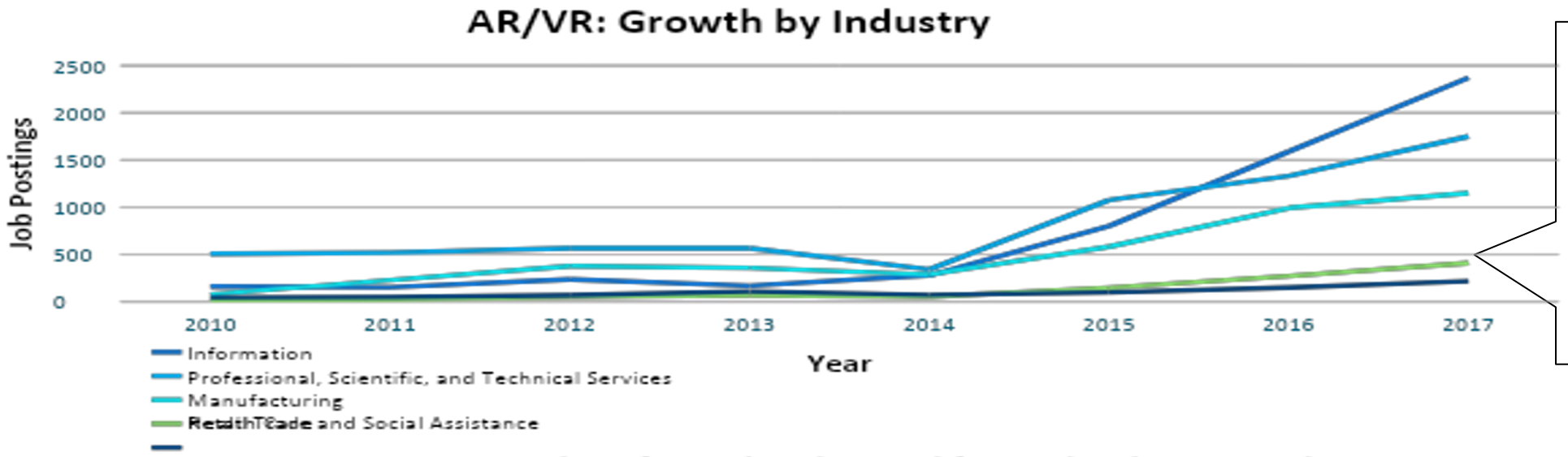
Identifying Future Skill Demands

A Range of Lenses for Tracking Emerging Trends

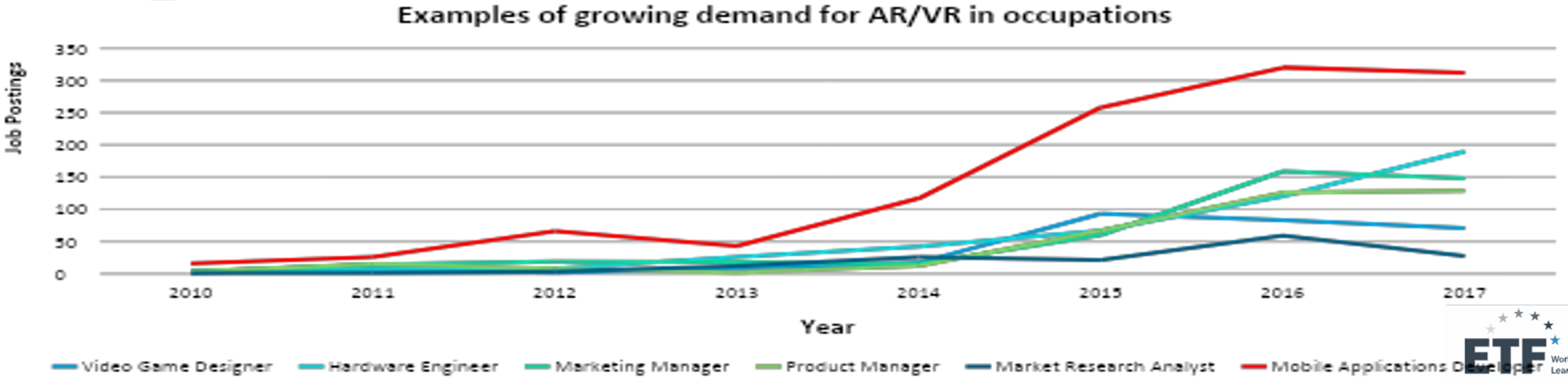
	Top IT Skills (Total postings)	Highest Paying IT Skills (Mean advertised salary)	Fastest Growing IT Skills (24 month projections)	Hardest to Fill IT Skills (Mean posting duration)
1.	SQL	Zookeeper	TensorFlow	Public Cloud Security
2.	Java	TensorFlow	General Data Protection Regulation (GDPR)	Infrastructure as a Service (IaaS)
3.	JavaScript	Scala	Kubernetes	Cloud Technology Architecture
4.	Linux	AWS Redshift	Spring Boot	Cloud Infrastructure
5.	Python	AWS DynamoDB	Webpack	Ansible
6.	Data Analytics	Go Programming Language (Golang)	AWS Lambda	Apache Mesos
7.	Salesforce	Pig	Salesforce Lightning	Data Protection Planning
8.	C#	Apache Mesos	Redux	Work Breakdown Structure
9.	Scrum	AWS CloudFormation	Financial Microservices	Hadoop Cloudera
10.	C++	Deep Learning	Apache Kafka	OpenShift

Evaluating skill demands of

Emerging Sectors



The retail industry, which to a large extent represents the consumers of AR/VR, has shown increasingly fast growth in demand after 2015

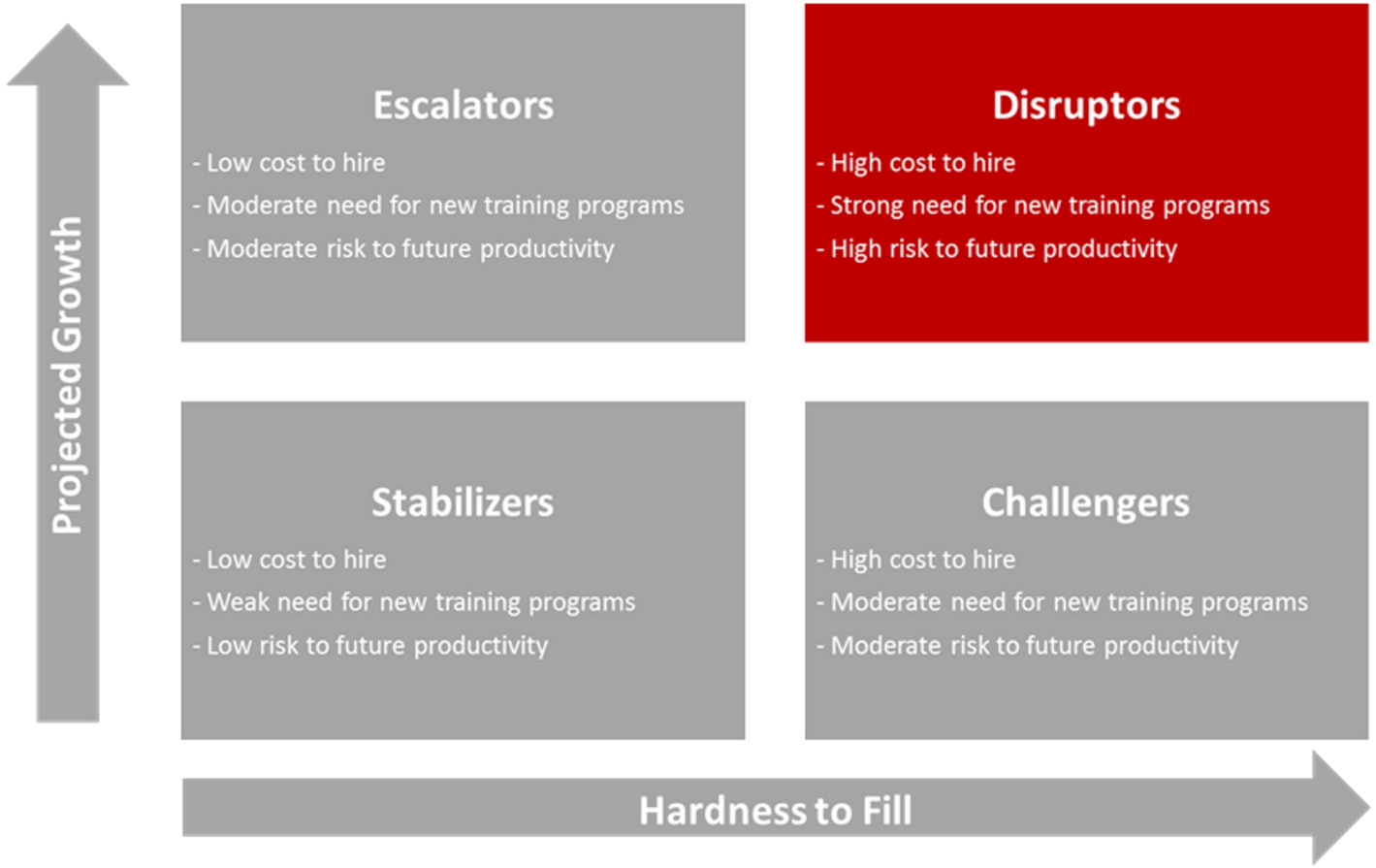


ETF Working together Learning for life
European Training Foundation

Identifying & Prioritizing

Disruptive Jobs & Skills

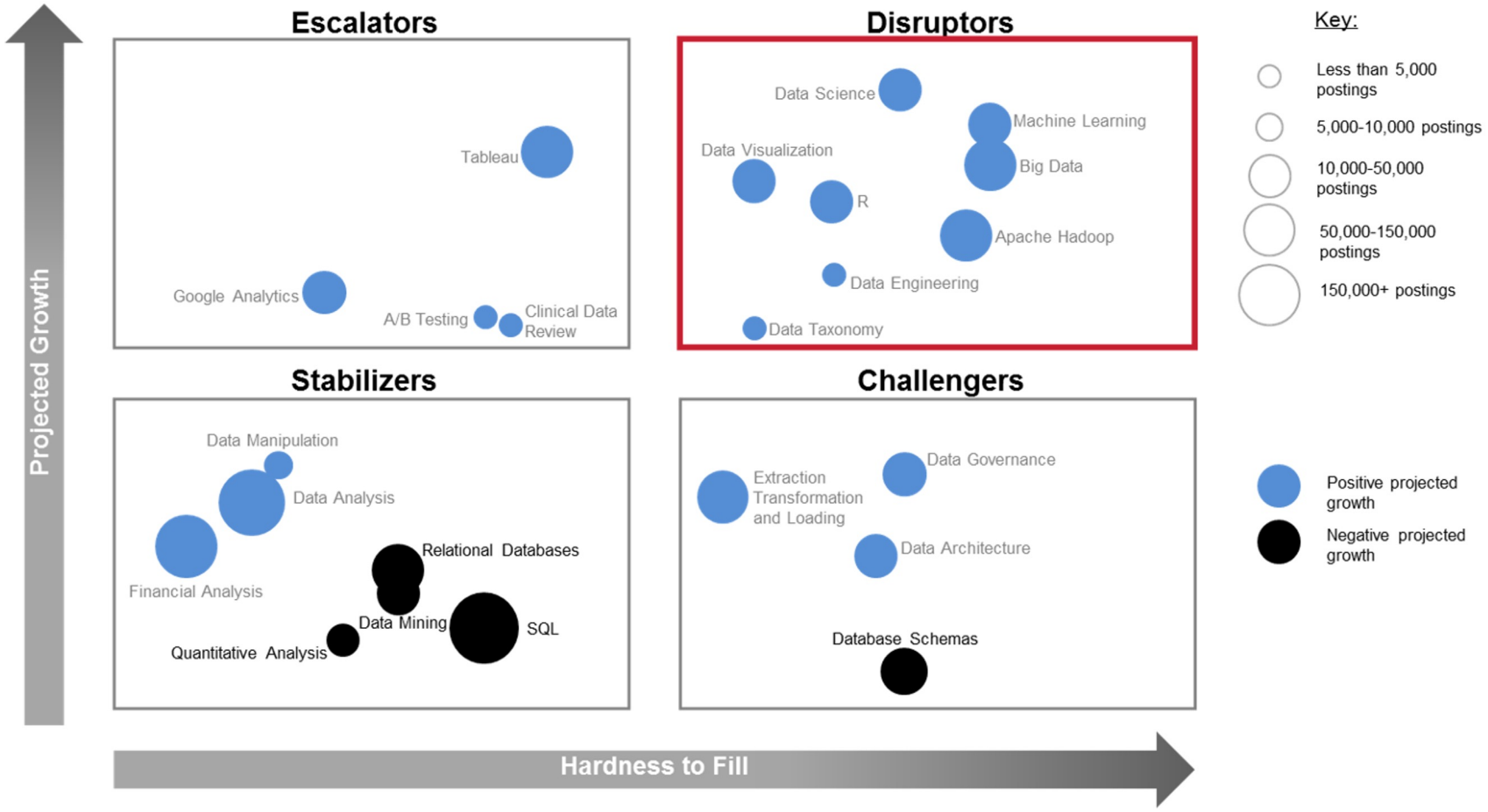
A framework for identifying and planning ahead for the skills that are likely to challenge the market into the future.



ETF Working together Learning for life
European Training Foundation

Even in Emerging Fields like Data Science

Not All Skills Are Created Equal

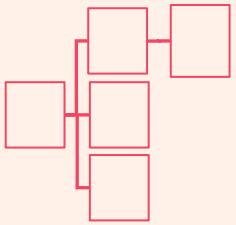


ETF Working together Learning for life
European Training Foundation

Connecting multiple sources

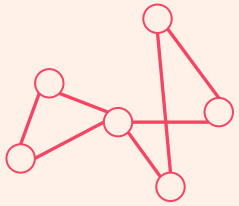
Lightcast

Structured data vs Big data



Structured data

Purposefully collected and collated data which comes in neat, tidy structure. In Emsi's case, this is typically data from government statistical surveys, designed to ask explicit questions of targeted samples of specific audiences.



Big data

Extremely large scale data captured from some transactions rather than as a specific data collection exercise. In Emsi's case, this means harvesting job postings and worker profiles from different web-based sources.

Aggregate data vs Micro data

Aggregate data

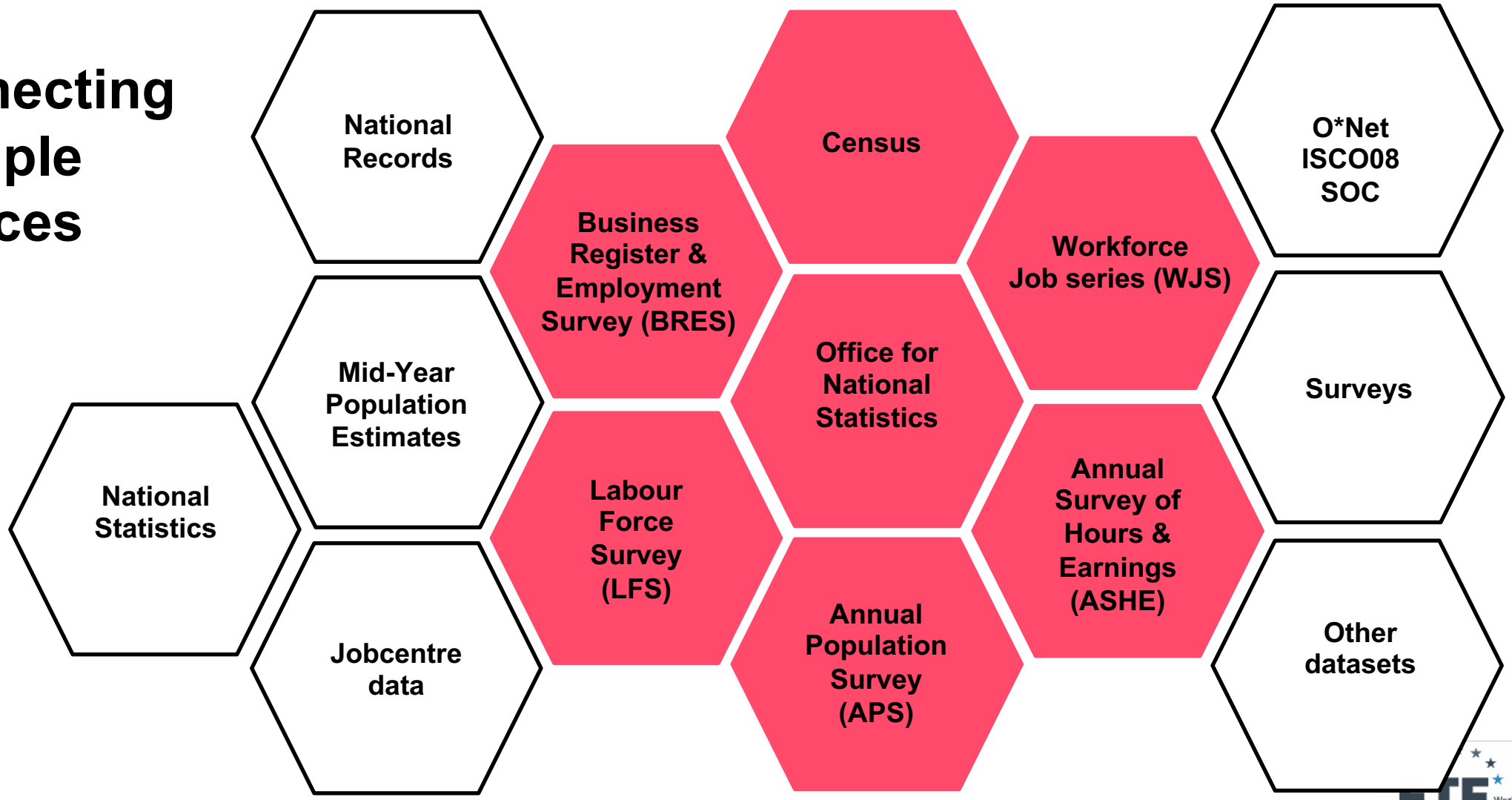
Aggregate data comes to us in the form of summary statistics: the total number of X, the average level of Y, the percentage doing Z. A lot of structured data is available only in aggregate form, with limited access to the micro data.

Micro data

Micro data are more powerful, with every individual case's data points available without aggregation; we can then cut it any way we like. Big data comes to us in micro data form, although we often prepare it for standard forms of aggregation.

Lightcast

Connecting multiple sources



ETF Working together Learning for life
European Training Foundation

Lightcast

Job Posting & Profile Analytics – document based LMI

- **JP & PA are Document-based LMI:** it's a Big Data source where we have the entire process in our control, including the raw data
- **JPA is (near) real-time:** we scrape job postings every day and within a day or two we can use the intelligence they produce
- **JPA is rich:** we have millions job ads (growing every day) from the last 2 years
- **JPA is noisy:** job postings can move differently to underlying labour market demand and so should be handled in that knowledge

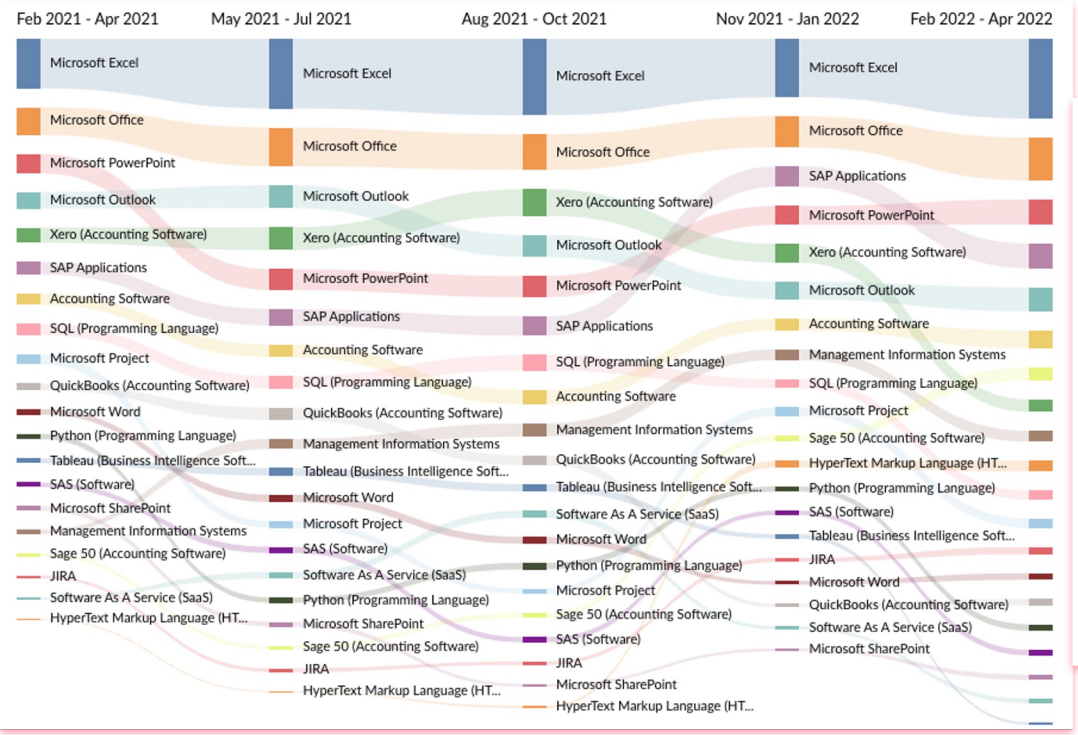
ETF Working together Learning for life
European Training Foundation

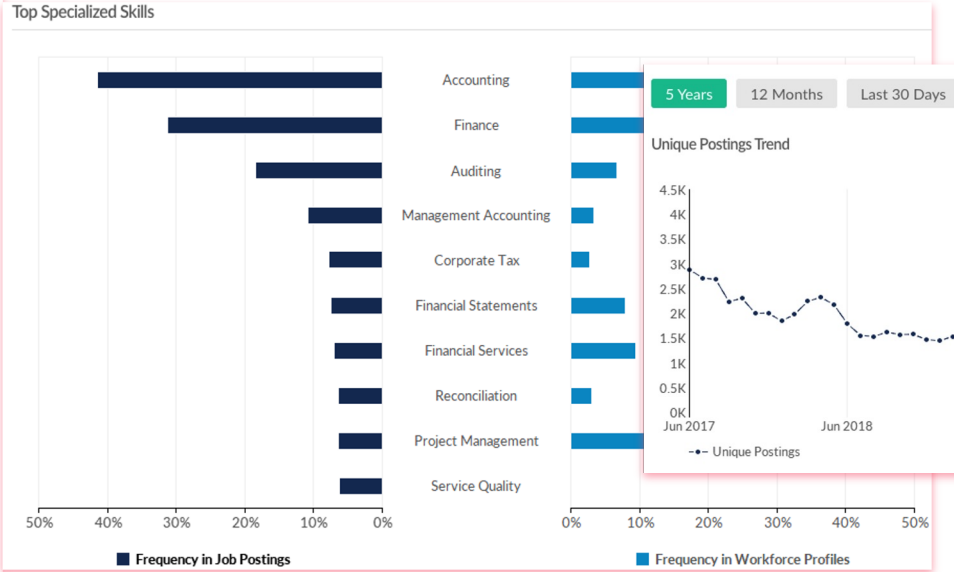
Live session

SkillsTrends | Skills Gaps | Recruitment Activity

from Lightcast Analyst (UK)

E.g. Occupations mapped to course area Accounting (JACS N400) in the City of Bristol, Bath and North East Somerset, North Somerset and South Gloucestershire April 2021 - May 2022





ETF Working together Learning for life
European Training Foundation

Multinational Skills Insights

from Lightcast Global

E.g. Analysts & Data Scientists with Python programming skills

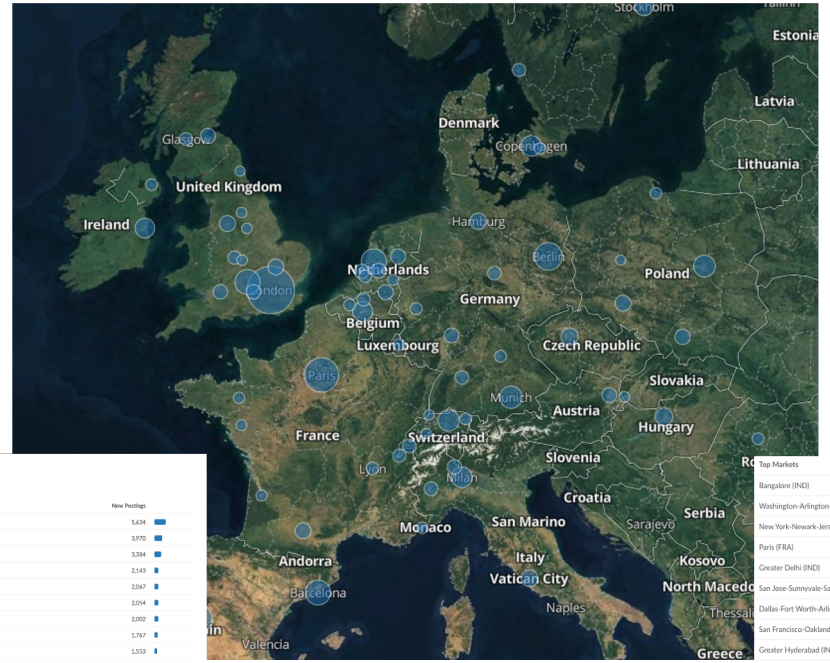
Identify concentrations of skills supply and demand across the world within occupations, and dig down into the data to glean insights from top skills for a particular population, top employers and population estimates.

From Apr 2021 to Mar 2022

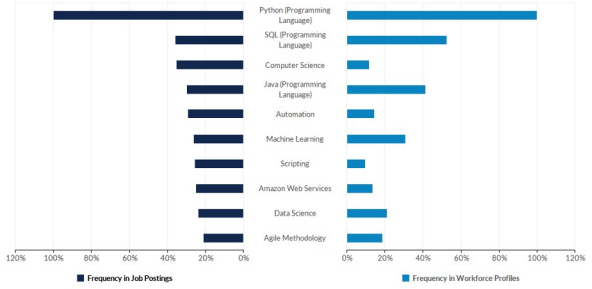
334,409
Unique Postings

35,317
Companies Posting

27%
Postings Increase



Top Specialized Skills



Top Companies

Companies with the most job postings from April 2021 to March 2022.

Top Companies	% Change	New Postings
Amazon	107%	5,634
Deloitte	637%	3,970
Humana	403%	3,384
Dell Technologies	321%	2,143
Facebook	591%	2,067
Accenture	236%	2,054
Verizon Communications	386%	2,002
Capital One	388%	1,767
Anthem Blue Cross	166%	1,553
KPMG	237%	1,537

Top Markets	% Change	New Postings
Bangalore (IND)	-27%	22,601
Washington-Arlington-Alexandria, DC-VA-MD-WV (USA)	11%	12,426
New York-Newark-Jersey City, NY-NJ-PA (USA)	47%	10,960
Paris (FRA)	-4%	8,510
Greater Delhi (IND)	-26%	8,334
San Jose-Sunnyvale-Santa Clara, CA (USA)	18%	8,115
Dallas-Fort Worth-Arlington, TX (USA)	74%	7,987
San Francisco-Oakland-Hayward, CA (USA)	5%	7,961
Greater Hyderabad (IND)	-38%	7,317
Berlin (DEU)	38%	7,296

Audit skills data from course documentation

from Lightcast Skill Sync

E.g. MSc Advanced Mechanical Engineering, QMUL, & BA Games Design & Art, Soton

Skill Sync

Simple Options | Advanced Options

Parse Text

H1 H2 | B I

On this MSc course, you can choose specialist pathways in Computational design & engineering or Energy & thermofluids.

You'll develop innovative computational and technological products for industry and perform a design analysis of mechanical engineering systems.

We'll build on your existing expertise by giving you specialised skills in energy storage engineering and robotics. You'll have the option to explore modules in computational engineering, renewable energy, numerical optimisation, heat transfer and fluid mechanics, materials selection, gas turbines, solar energy, combustion, nanocomposites and manufacturing process.

You'll also complete a substantial research project in our excellent facilities, supervised by our research staff. These could feature a wide range of internal combustion engines, heat transfer facilities, wind tunnels, a flight simulator, an induction jet engine test bench, and materials synthesis and characterisation labs. If your interests lie in nanotechnology, you can call on the facilities and expertise provided by Nanoforce, a company directly associated with the School.

1,115 / 10,000 allowed characters Clear

Suggested Job Titles Create Job Titles Group

HVAC Systems Engineers | Engineering Department Interns | Research Hydrologists | Process Design Engineers | Aerodynamics Engineers

Filter profiles by occupations Clear · Save Browse · Groups

Mechanical Engineers (2122) ✖

Search Logic

Profiles with all of the selected skills

Skills	% Profiles
Mechanical Engineering	18%
Research	3%
Innovation	2%
Turbines	1%
Fluid Mechanics	0.5%
Manufacturing Processes	0.4%
Material Selection	0.2%
Wind Tunnel	0.1%
Heat Transfer	0.1%
Internal Combustion Engines	0.1%
Nanotechnology	0.04%

Create Group From Selected Skills

Skill Sync

Simple Options | Advanced Options

Parse Text

H1 H2 | B I

You'll learn to apply design-thinking to the creation of brand new games and refine them through research-led creative investigation, making, prototyping and user testing

You'll explore the full spectrum of gaming, including screen-based (video) games, augmented and virtual reality game-making, alongside the physical production of products such as location-based games, board games, card games and escape rooms. You'll also cover the emerging fields of purposeful games and serious games, and explore how games can be designed to have social impact and cultural significance.

Throughout your studies you'll work with your peers in a small, close-knit studio environment. You'll collaborate and inspire one another, and discover your strengths. You'll gain a range of skills including illustration, character design and storytelling as well as coding and digital production.

Your course includes:

pitching your ideas and prototypes to industry

visits to innovative games studios such as Hello Games, Sensible Object, ustwo games and Sennep games

visits to game conferences

workshops and advice on your projects from industry

live projects with progressive organisations looking to utilise game theory and game design such as National Air Traffic Control, IBM and the Royal Society of Public Health

2,004 / 10,000 allowed characters Clear

Suggested Job Titles Create Job Titles Group

UX Designers | UI/UX Designers | Product Designers | Lead UX Designers | Graphic Designers

Filter profiles by occupations Clear · Save Browse · Groups

Programmers and Software Developm... ✖

Web Design and Development Profess... ✖

Profiles with all of the selected skills

Skills	% Profiles
Research	7%
Innovation	3%
Prototyping	1%
Animations	1%
Game Design	1.0%
Illustration	0.8%
Usability Testing	0.6%
Virtual Reality	0.4%
Design Thinking	0.2%
Air Traffic Control	0.1%
Creative Industries	0.1%

Create Group From Selected Skills

Start Over

Timeframe

Aug 2018 to Sep 2022

1 Occupation Advanced · Clear · Save

Results should...

Include

✖ Programmers and Software Developmen...

Add Occupations...

1 Nation Advanced · Clear · Save

✖ United Kingdom (K02000001)

Add Regions...

Job Titles

None selected.

Add Job Titles...

Companies Groups

None selected.

Add Companies...

Include Skills & Qualifications

Has ANY of the following skills

Groups

None selected.

Add Skills & Qualifications...

- AND -

Has ALL of the following skills

Groups

None selected.

Add Skills & Qualifications...

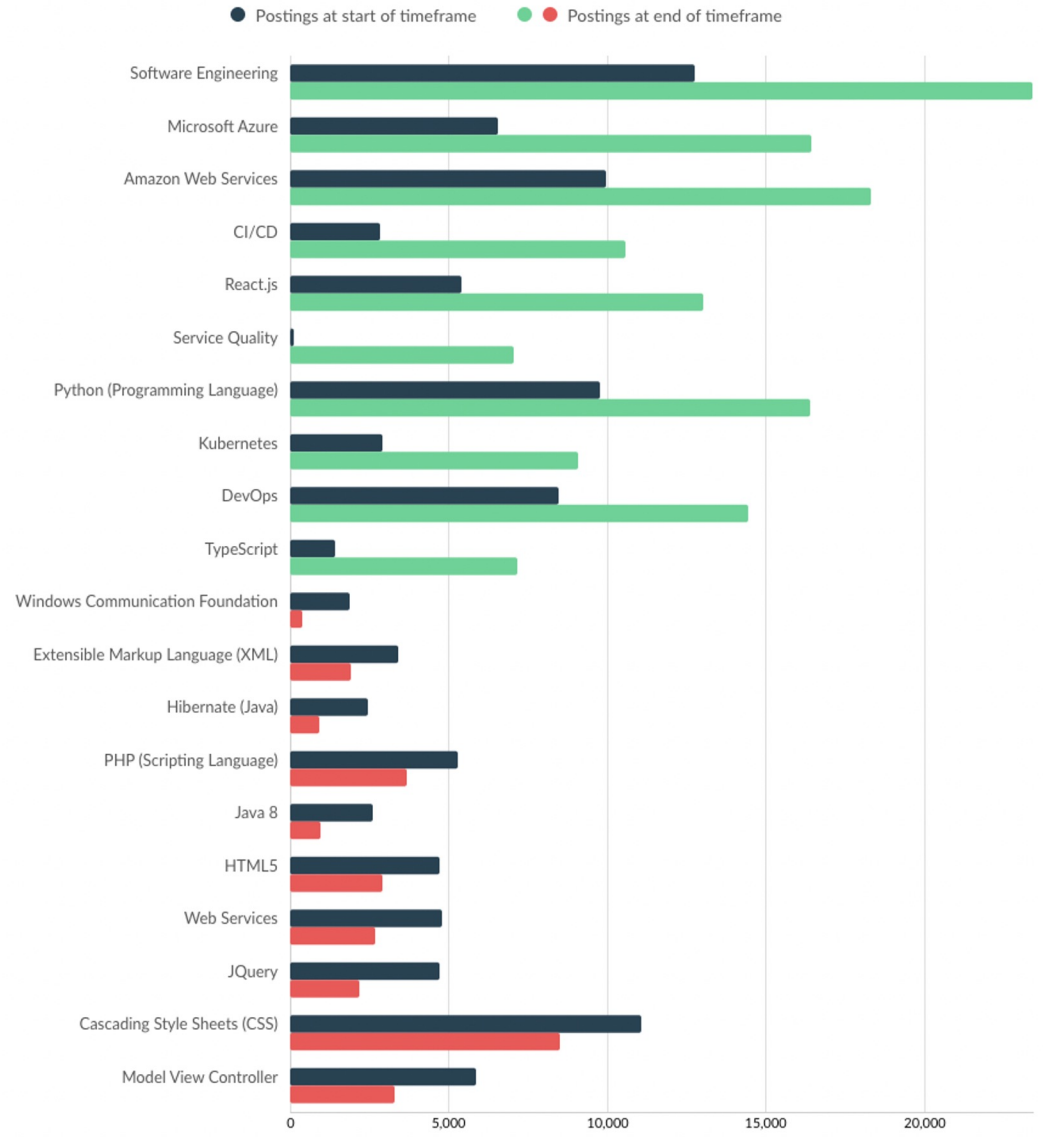
Exclude Skills & Qualifications Groups

Results should...

Exclude any of the selected skills

Hot and Cold Skills by Job Postings

Save Export



Skill	Postings in Aug 2018	Postings in Sep 2022	Change	Percent Change

ETF Working together Learning for life

European Training Foundation