Big Data for Labour Market Intelligence

Programme de développement des capacités 2022

Module 1 : Formation technique

Session 9

Utilisation des Big Data pour l'analyse des dynamiques professionnelles et des compétences sur le marché de l'emploi
Combinaison de données issues de sources différentes

Intervenant : Mauro Pelucchi

29/11/2022

Système de production des données

Tunisie (avril 2020-juillet 2022)

- ✓ 680 191 OEL - > 175 203 dédupliquées

Ukraine - général (avril 2020-juillet 2022)

- ✓ 2 571 655 OEL - > 1 304 262 dédupliquées

Géorgie (avril 2021-juillet 2022)

- ✓ 129 271 OEL - 84 817 dédupliquées

Égypte (nouveau)

- ✓ 1 307 678 OEL - 391 701 dédupliquées

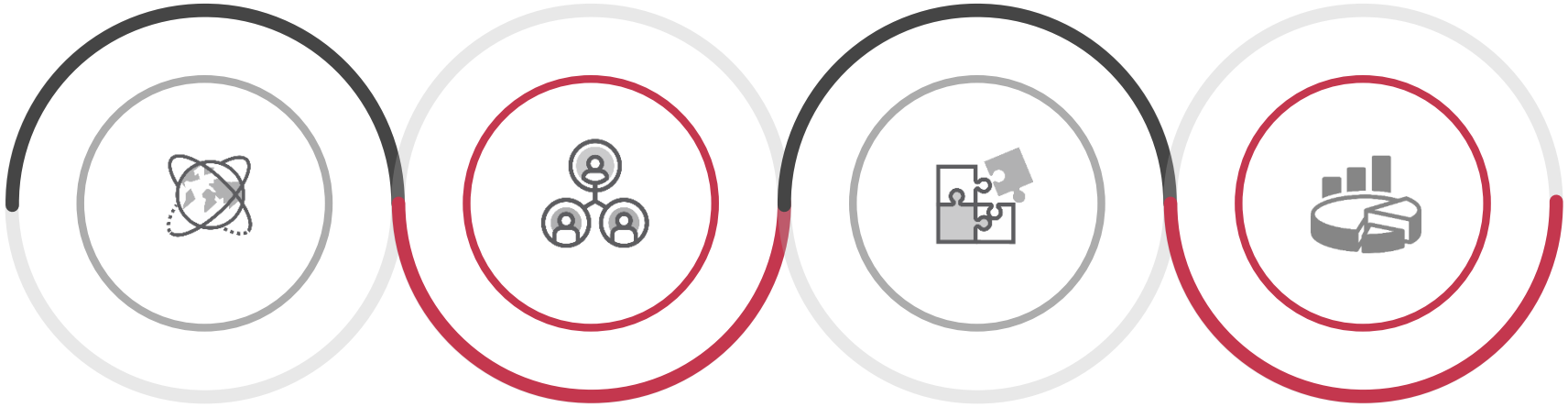
Kenya (nouveau)

- ✓ (la collecte a commencé en septembre 2022)

ETF Working together Learning for life
European Training Foundation

Collecte et décodage des données sur le marché du travail

Les données sur le marché de l'emploi en temps réel fournissent des informations actualisées que les sources traditionnelles ne permettent pas d'obtenir.



Saisie de données sur le marché du travail

Marquage et structuration

Langue commune

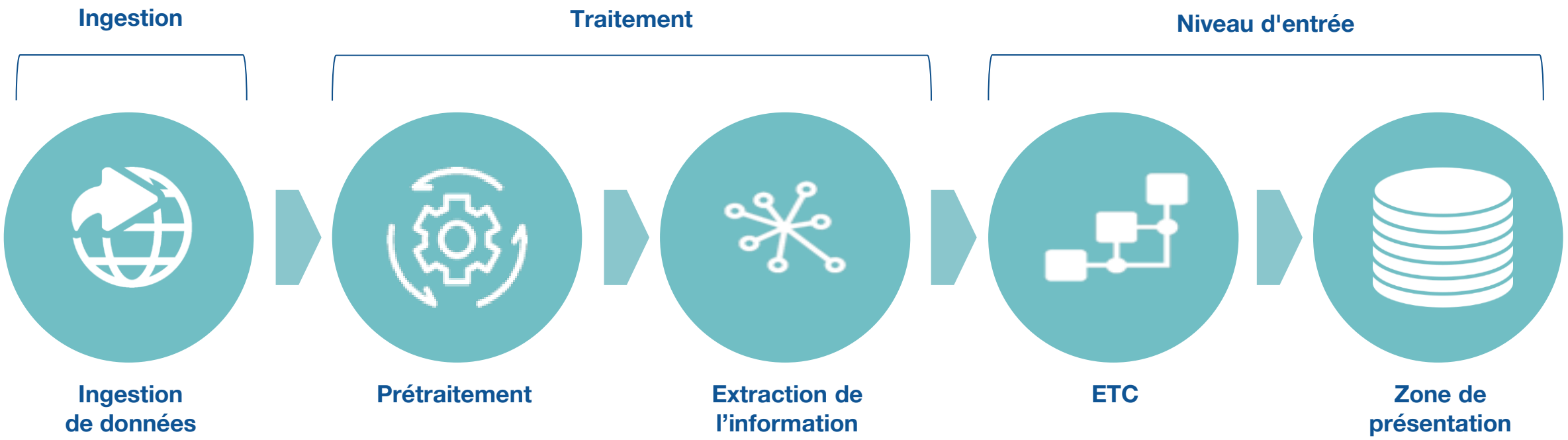
L'ontologie des données permet de faire des comparaisons.

Formulation de conclusions

Indication sur les compétences recherchées et les tendances des carrière dans la vie réelle.

ETF Working together Learning for life
European Training Foundation

Flux de données global



Ingestion de données

- **Objectif :**
  - Collecte d'une quantité considérable de données hétérogènes provenant d'un nombre imprévisible de sources.
- **Défis :**
  - Garantir l'exhaustivité et la cohérence des données.
- **Approche :**
  - Développer un cadre multitechnique (crawling, scraping, API) pour faire correspondre les différentes caractéristiques des sites Web : volume d'offres d'emploi, langues utilisées, structure technologique, approche non invasive et accords politiques.
  - Prévenir les pertes de sources de données par le biais de politiques de redondance.
  - Détecter et collecter des métadonnées pour améliorer la valeur des informations.
- **Caractéristiques :**
  - Pour garantir la gouvernance du processus dans la phase d'ingestion des données, nous avons développé un système de surveillance avec un outil de planification et un module d'alerte.

Prétraitement des données

- **Objectif**
  - Alimenter la phase d'extraction des informations avec les bonnes données.
- **Défis**
  - Mesurer, surveiller et améliorer la qualité des données, pour en maximiser l'exhaustivité, la cohérence, la complexité, l'actualité et la périodicité.
- **Approche**
  - Développer un pipeline à plusieurs phases, axé sur :
    - la détection des offres d'emploi : analyse d'une page Web pour sélectionner uniquement le contenu relatif aux offres d'emploi.
    - la déduplication : détecter les offres d'emploi en double pour obtenir une seule entité d'offre d'emploi.
    - la détection des date : identifier les dates de publication et d'expiration grâce à l'analyse de la description des offres d'emploi.
    - la durée de l'offre : méthode pour définir la date d'expiration, lorsqu'elle n'est pas explicitement disponible.
- **Caractéristiques** :
  - Garantir la qualité des données pendant toutes les phases du traitement.

ETF Working together Learning for life
European Training Foundation

Classification des données

- **Objectif :**
  - **Extraire et structurer les informations des données pour les fournir au niveau présentation.**
- **Défis :**
  - **Traiter des quantités considérables de données hétérogènes écrites dans des langues différentes.**
- **Approche :**
  - **Développer un cadre adaptable, dépendant de la langue, adapté aux différentes caractéristiques de l'information. Quelques défis pertinents :**
    - **Classification des caractéristiques de Profession : méthodes combinées comme l'apprentissage automatique, la modélisation thématique et l'apprentissage non supervisé.**
    - **Classification des caractéristiques de Compétences : autres méthodes combinées, comme l'analyse de texte avec similarité du corpus ou des connaissances.**
- **Caractéristiques :**
  - **Garantir l'extraction d'informations explicables, les méthodes de classification des collectes et les caractéristiques pertinentes.**

Présentation

- **Objectif :**
  - Fournir des données dans un environnement analytique pour répondre aux différents besoins d'analyse des utilisateurs.
- **Défis :**
  - Accès personnalisé aux données pour les différents types de parties prenantes.
- **Approche :**
  - Développer différents chemins de navigation, pour différents utilisateurs :
    - **Analystes de données/experts en mégadonnées** : accès via un environnement de laboratoire de découvertes pour répondre à leur besoin d'analyser des données libres.
    - **Décideurs** : accès aux informations par le biais de mises en récit et de tableaux de bord dynamiques.

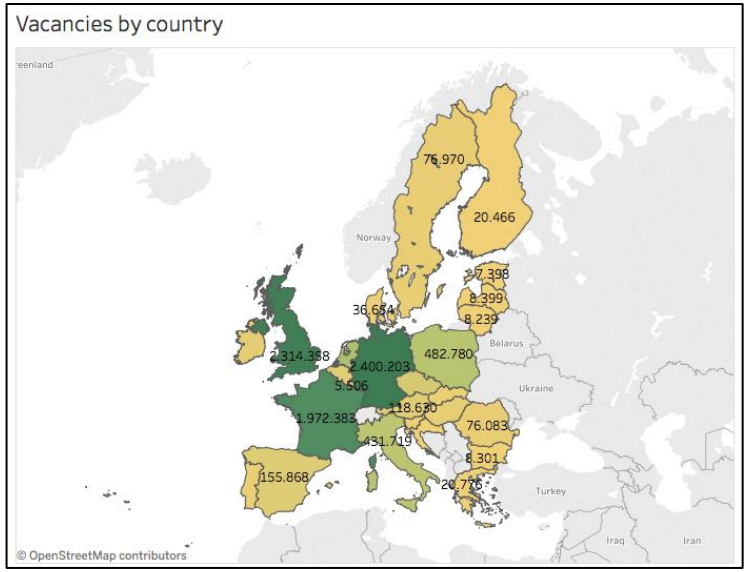
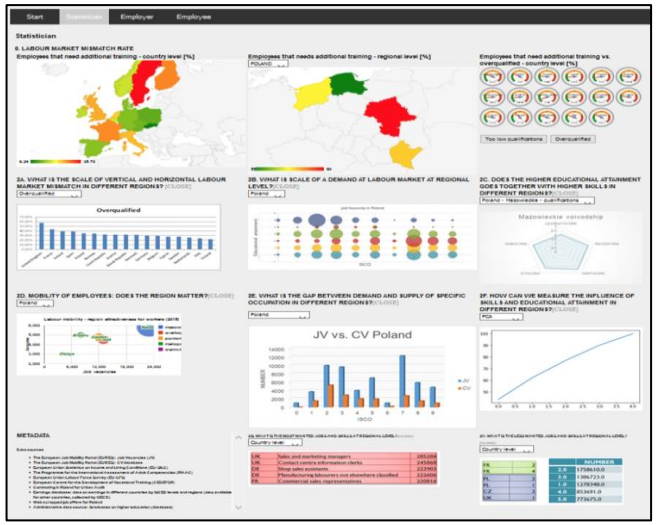


Tableau de bord
avec possibilité de
pointer-cliquer



Portail du marché du travail
(extrait de Big Data Hackathon 2017 - équipe Pologne)

Carte (extrait de Big Data Hackathon 2017 - équipe France)

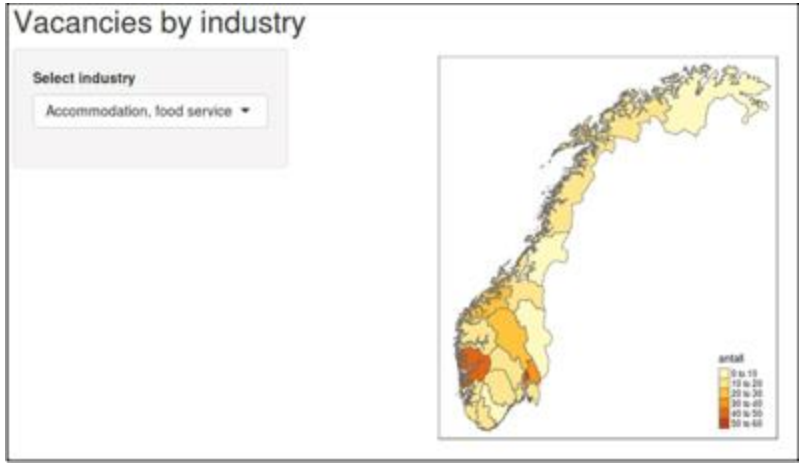


Tableau de bord (extrait de Big Data Hackathon 2017 - équipe Norvège)

Session de laboratoire
Avantage comparatif révélé (ACR)

Lightcast

ACR

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:

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

ETF Working together Learning for life
European Training Foundation

Similarité de la profession

Du bloc-notes :

- Nous devons sélectionner un pays.
- Pour le pays, nous devons calculer l'ACR
- Ensuite nous pourrons voir le pont créé entre les professions.

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
```

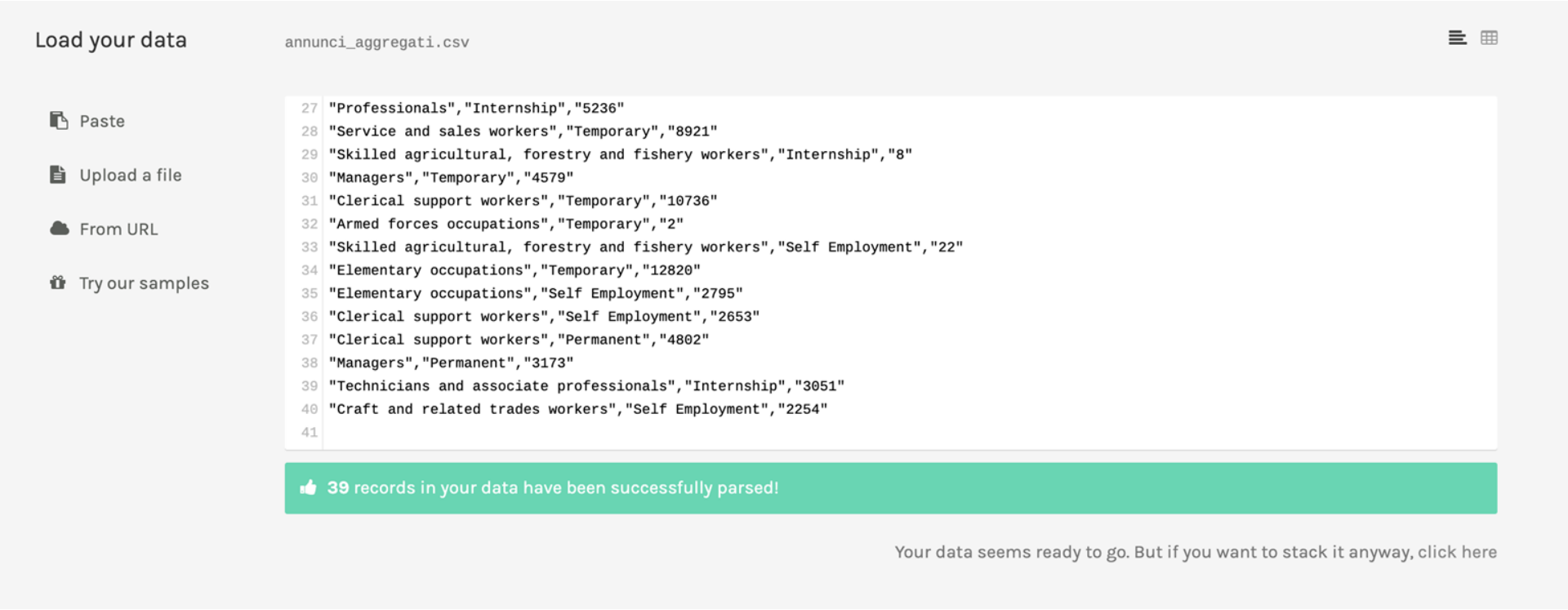
(1) Spark Jobs

	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.

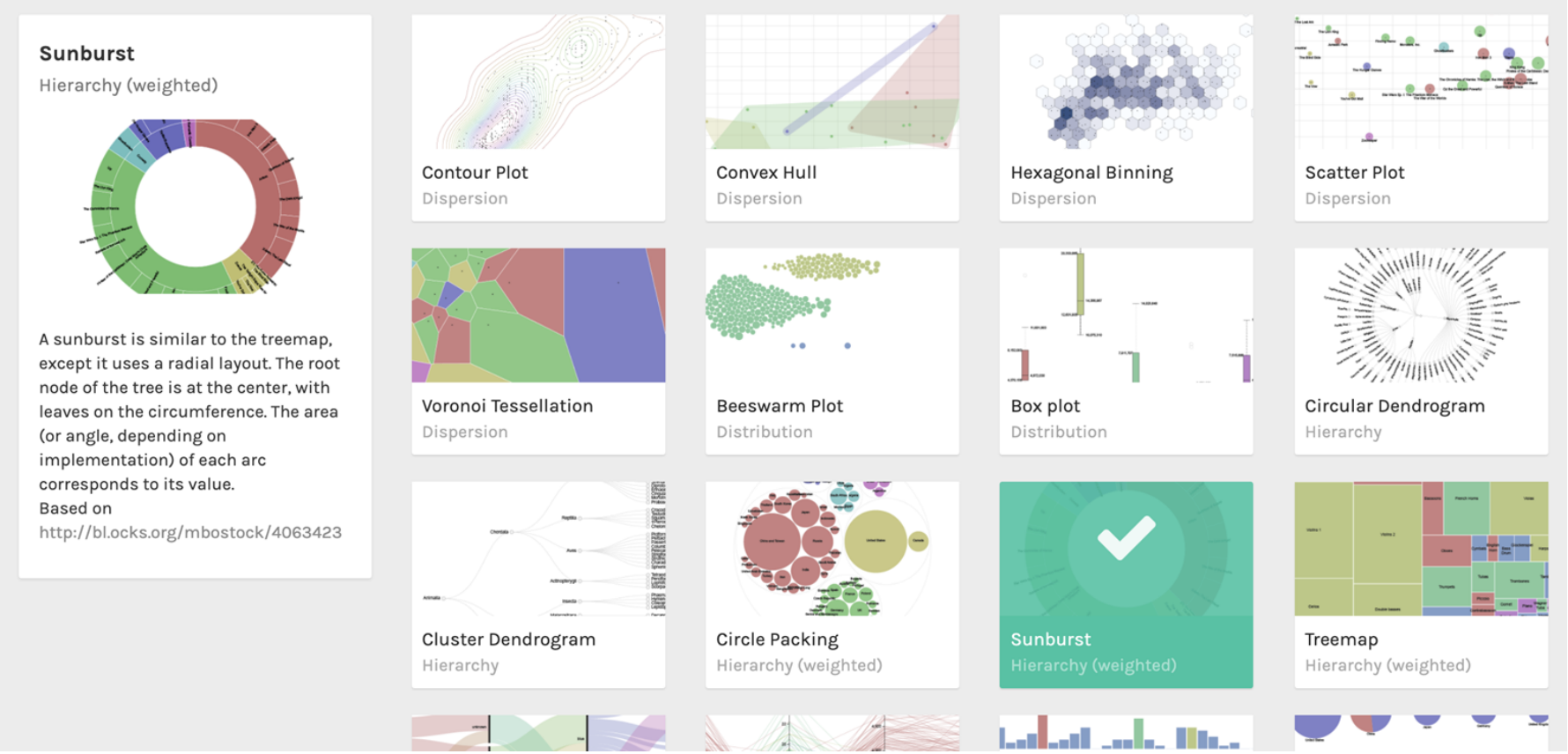
rawdata.io

https://app.rawgraphs.io/



rawdata.io

https://app.rawgraphs.io/

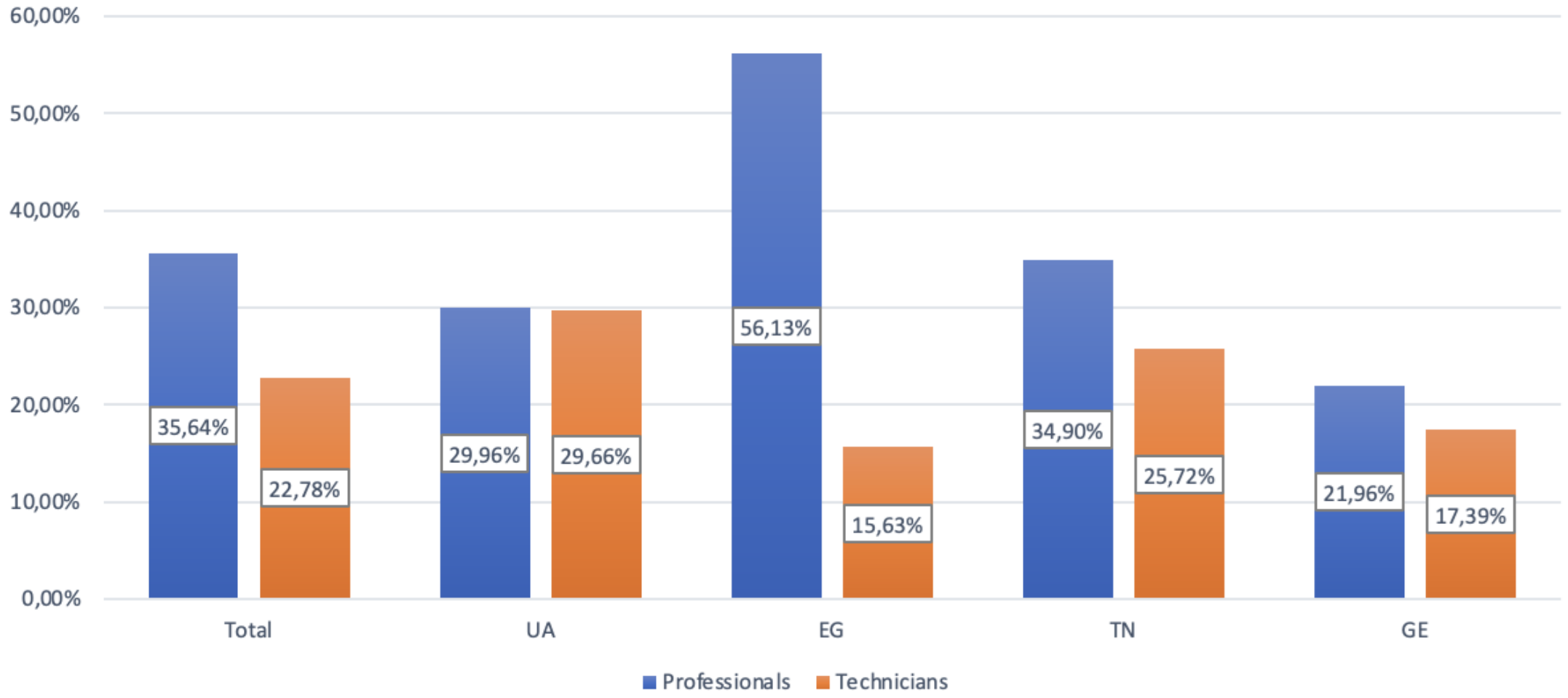


ETF Working together Learning for life
European Training Foundation

Compétences et professions

Lightcast

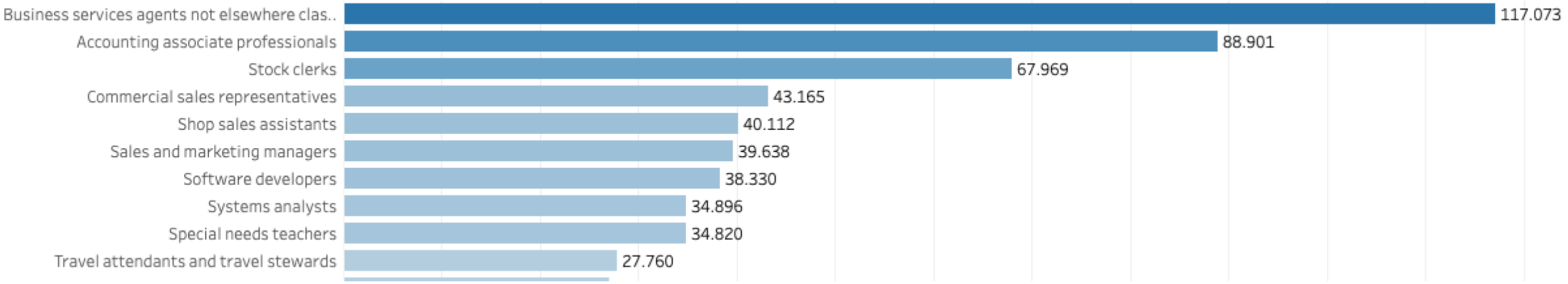
Un rôle important pour les professionnels



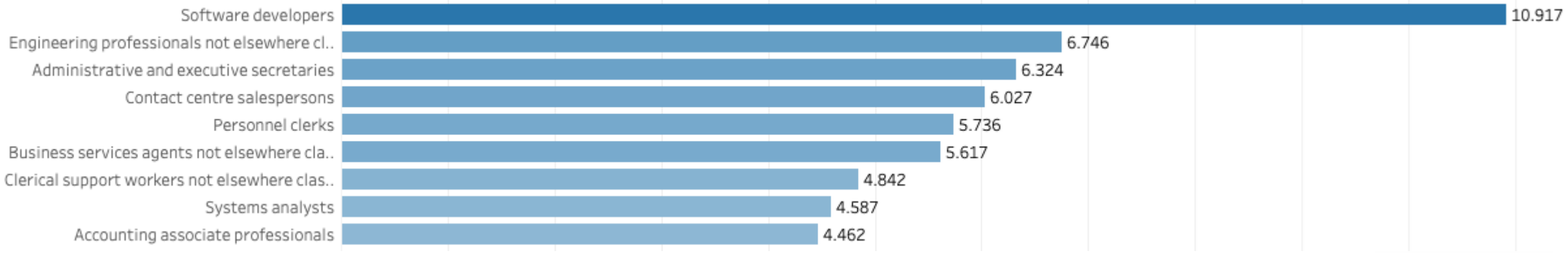
Source : Jeu de données ETF - sept. 2020-2022.

Professions détaillées

UA - Jeu de données ETF - Offres d'emploi uniques 2020-2022



TN - Jeu de données ETF - Offres d'emploi uniques 2020-2022



Compétences et ESCO

La taxonomie ESCO permet de regrouper les compétences en catégories. Distinction importante à faire entre compétences techniques et compétences non techniques :

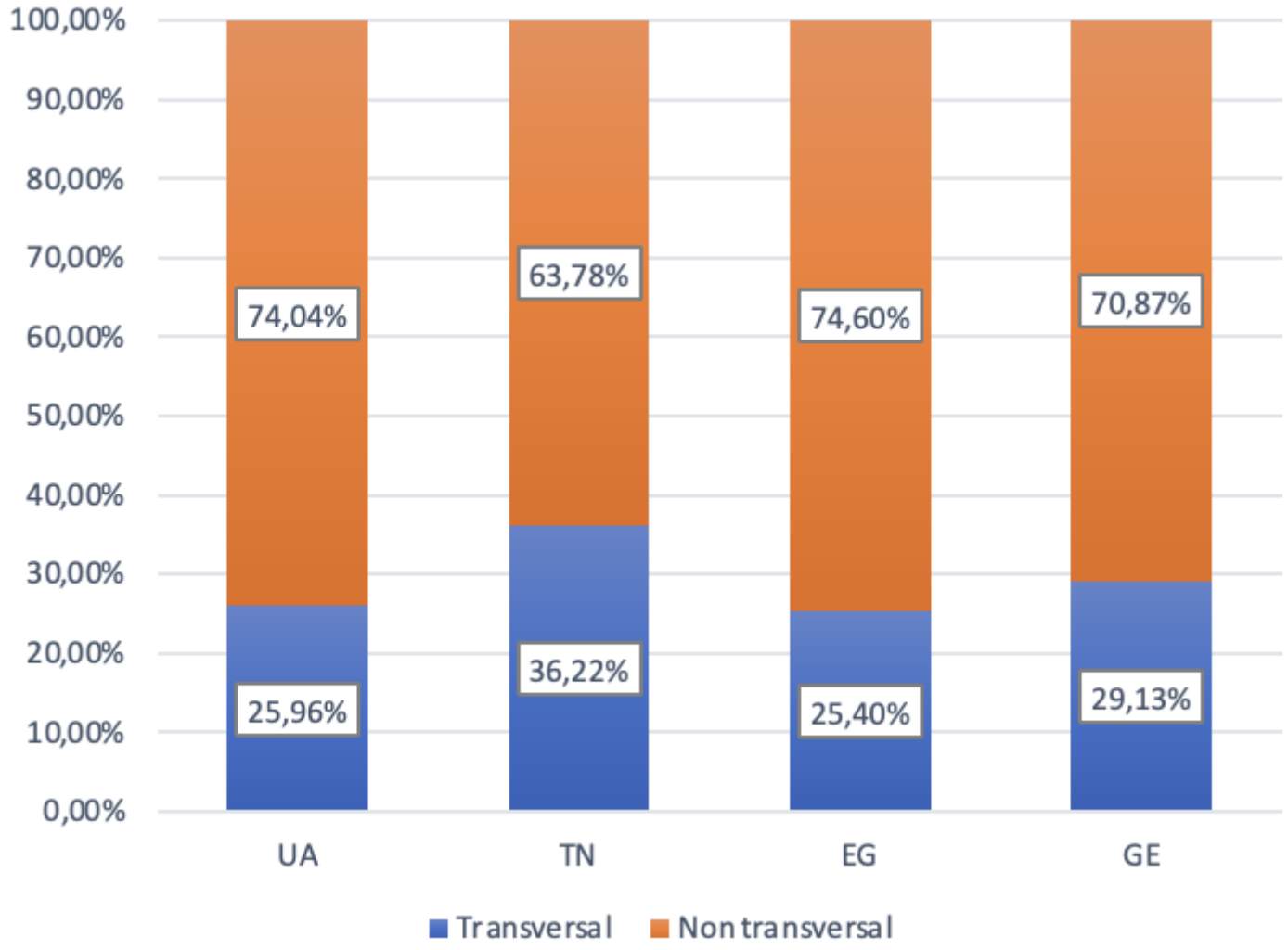
- **Compétences techniques** : compétences et aptitudes spécifiques au travail, nécessaires pour réaliser une tâche ou un travail spécifique (ex. : connaissances de logiciels ou d'instruments spécifiques, capacités manuelles spécifiques, etc.).
- **Compétences non techniques** : transversales par nature, elles désignent la capacité des individus à interagir avec les autres et avec l'environnement (ex. : capacité de communication, de résolution de problèmes, etc.).

Compétences transversales de l’ESCO

- **Compétences en matière de raisonnement :** capacité d'appliquer des processus mentaux et de raisonner pour résoudre des problèmes complexes, accroître les connaissances et effectuer des tâches complexes.
- **Interaction sociale :** capacité à interagir et collaborer avec les collègues, les clients et les consommateurs.
- **Application des connaissances :** application générale des compétences couramment utilisées sur le lieu de travail et en apprentissage ; connaissance de l'organisation et de l'environnement de travail.
- **Aptitudes et valeurs :** style de travail de la personne, ses préférences et ses opinions sur le travail qui définissent son comportement. Par exemple, la capacité à s'adapter au changement, à travailler de manière indépendante, à respecter les engagements, etc.

Nous ajoutons les langues aux compétences transversales.

Compétences : transversales et non transversales



ETF Working together Learning for life
European Training Foundation

Compétences transversales les plus pertinentes - UA

Compétences non transversales les plus pertinentes - UA

Marché du travail
Facteurs de défi

Lightcast

Marché du travail – Facteurs de défi

- Évolution des compétences
- Nouvelles professions émergentes
- Automatisation/remplacement des tâches
- …

Calcul du taux et de la pertinence des compétences

Le **taux de compétences** estime la part des compétences numériques, techniques et non techniques dans une seule profession.
La **pertinence des compétences** définit le poids de chaque compétence pour l'emploi en question.

Idée : Exploiter le pouvoir informatif de l'OEL pour calculer le taux de compétence et la pertinence des compétences.

La méthodologie du taux et de la pertinence des compétences a été définie et utilisée par CRISP en 2017 et 2018 pour l'Observatoire italien des compétences numériques promu par les associations italiennes de TIC.

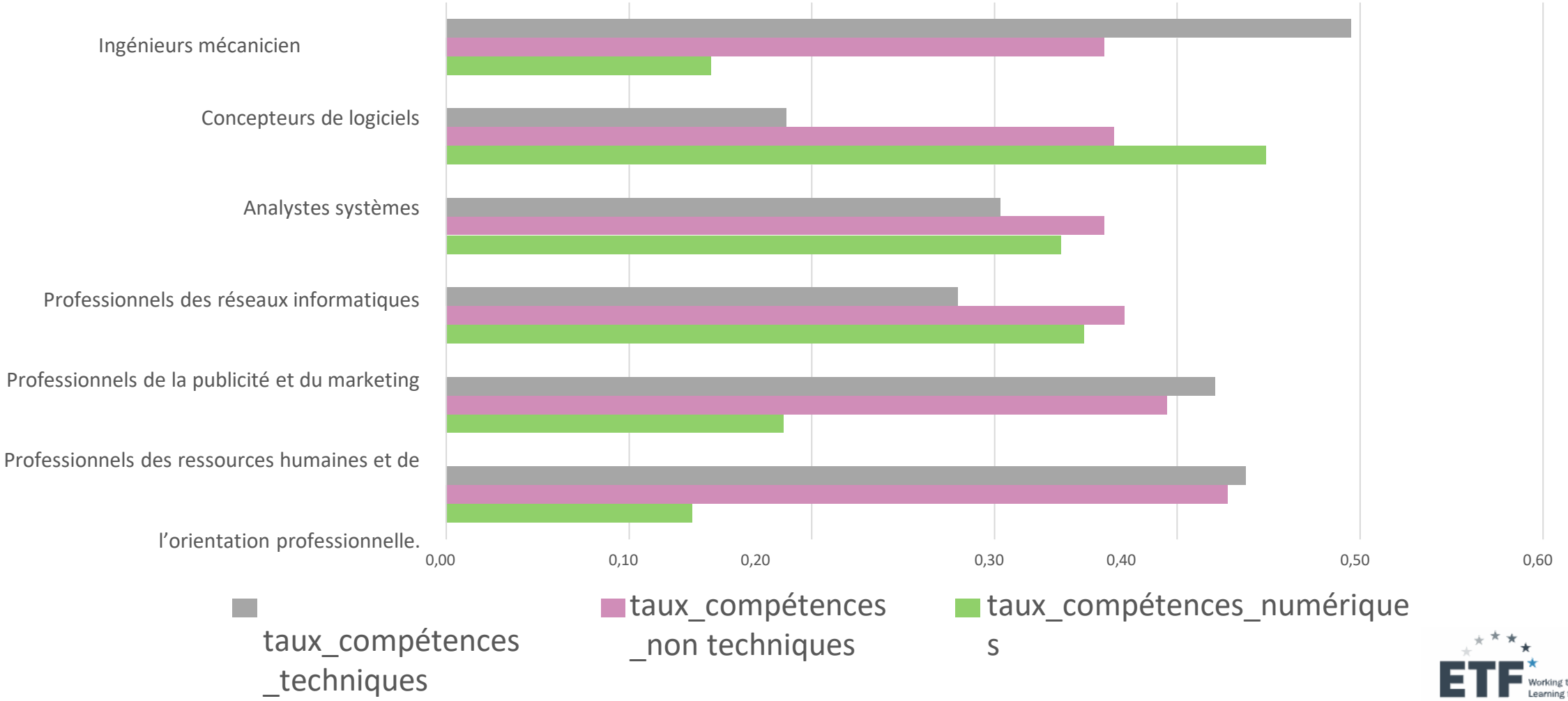
Working together
Learning for life
Training Foundation

Définition et méthode de calcul

$$\textit{Taux de compétences} = \frac{\textit{fréquence des compétences (numériques ou techniques ou non techniques)}}{\textit{Fréquence des compétences numériques + techniques + non techniques}}$$

$$\text{Pertinence des compétences} = \frac{\textit{n° d'offres d'emploi de (compétences , emploi)}}{n^\circ \text{ d'offres d'emploi de (compétence)}} \times \frac{n^\circ \text{ d'offres d'emploi de (emploi, compétence)}}{n^\circ \text{ d'offres d'emploi de (empoi)}}$$

Taux de compétences pour certaines professions



Observations définitives

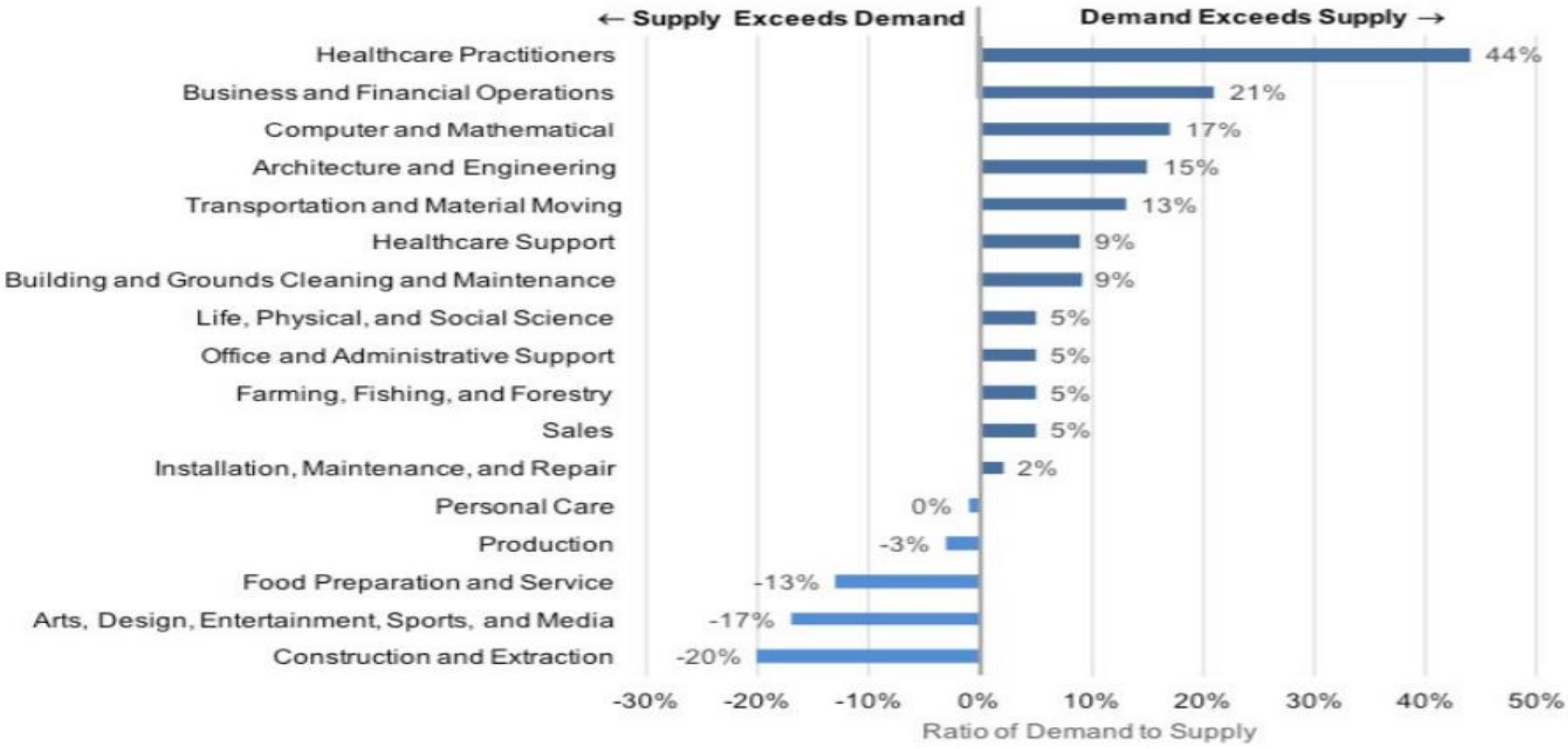
- Les connaissances sont un élément essentiel de l’observation des facteurs posant un défi pour le marché du travail :
  - ✓ Évolution des compétences
  - ✓ Nouvelles professions émergentes
  - ✓ Automatisation/remplacement des tâches
  - ✓ …
- Les OEL sont complémentaires des autres sources que nous utilisons normalement pour comprendre les phénomènes du marché du travail.
- Nous ne sommes qu'au début de l'exploitation du potentiel informatif des OEL-Big Data.

Informations issues des Big Data

Lightcast

Tous les manques ne sont pas des déficits

Caractérisation plus approfondie du déficit de compétences



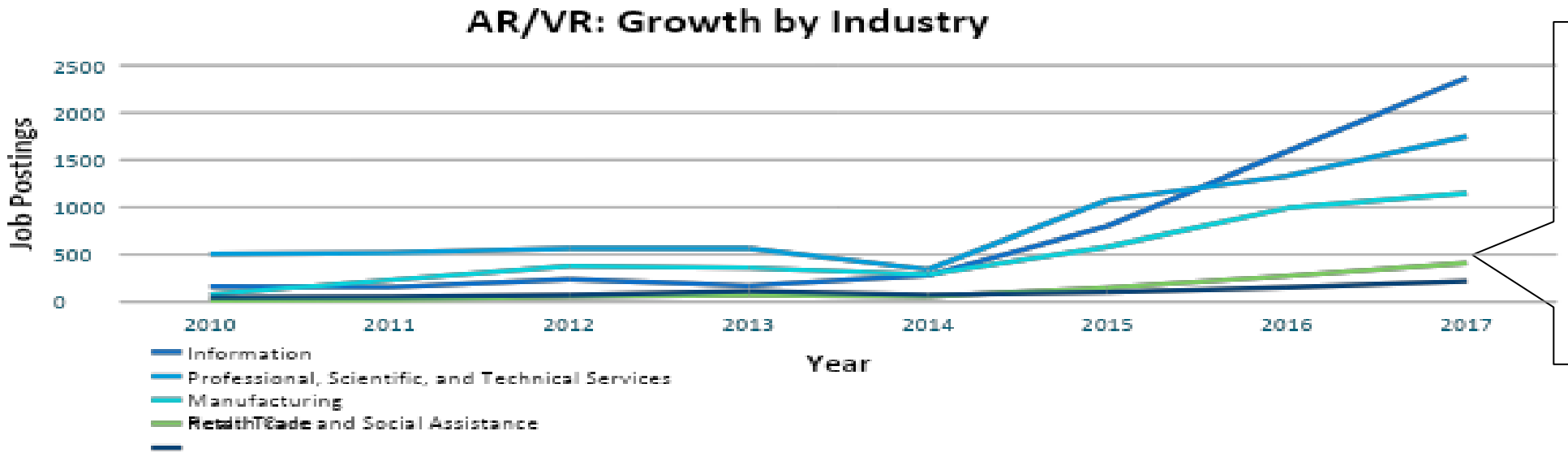
European Training Foundation

Identification des futures exigences en matière de compétences

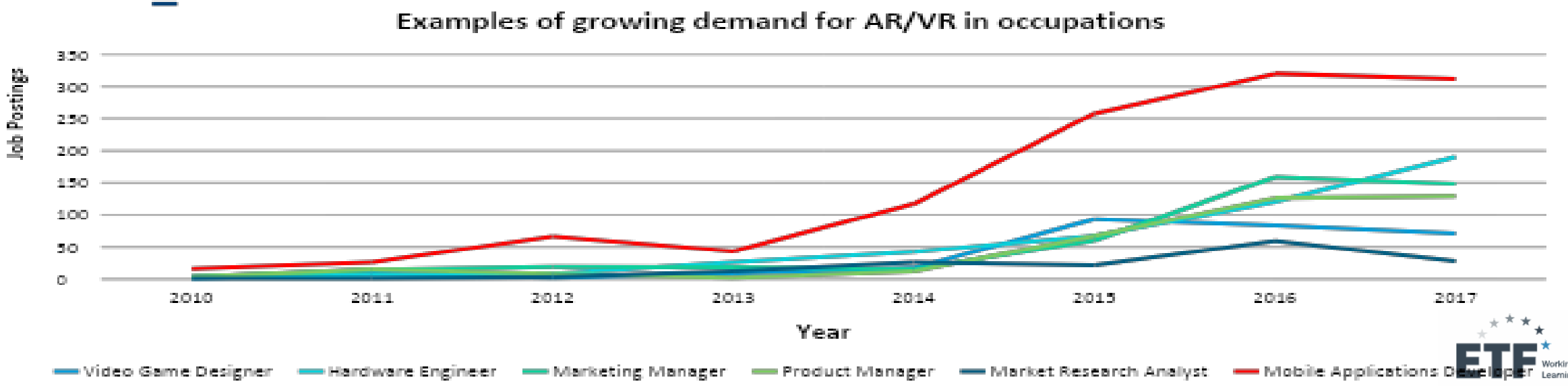
Plusieurs angles de suivi des tendances émergentes

	Principales compétences informatiques (total des offres publiées)	Compétences informatiques les mieux rémunérées (salaire moyen publié)	Compétences en informatique avec la croissance la plus rapide (projections sur 24 mois)	Compétences informatiques les plus difficiles à pourvoir (durée moyenne de publication)
1.	SQL	Zookeeper	TensorFlow	Sécurité du cloud public
2.	Java	TensorFlow	Règlement général sur la protection des données (RGPD)	Infrastructure à la demande (IaaS)
3.	JavaScript	Scala	Kubernetes	Architecture de la technologie cloud
4.	Linux	AWS Redshift	Spring Boot	Infrastructure cloud
5.	Python	AWS DynamoDB	Webpack	Ansible
6.	Analyse de données	Langage de programmation Go (Golang)	AWS Lambda	Apache Mesos
7.	Salesforce	Pig	Salesforce Lightning	Planification de la protection des données
8.	C#	Apache Mesos	Redux	Structure de répartition du travail
9.	Scrum	AWS CloudFormation	Microservices financiers	Hadoop Cloudera
10.	C++	Apprentissage profond	Apache Kafka	OpenShift

Évaluation des exigences en matière de compétences
secteurs émergents



La demande dans le secteur du commerce de détail, représentant notamment les consommateurs de RA/RV, a fortement augmenté depuis 2015.

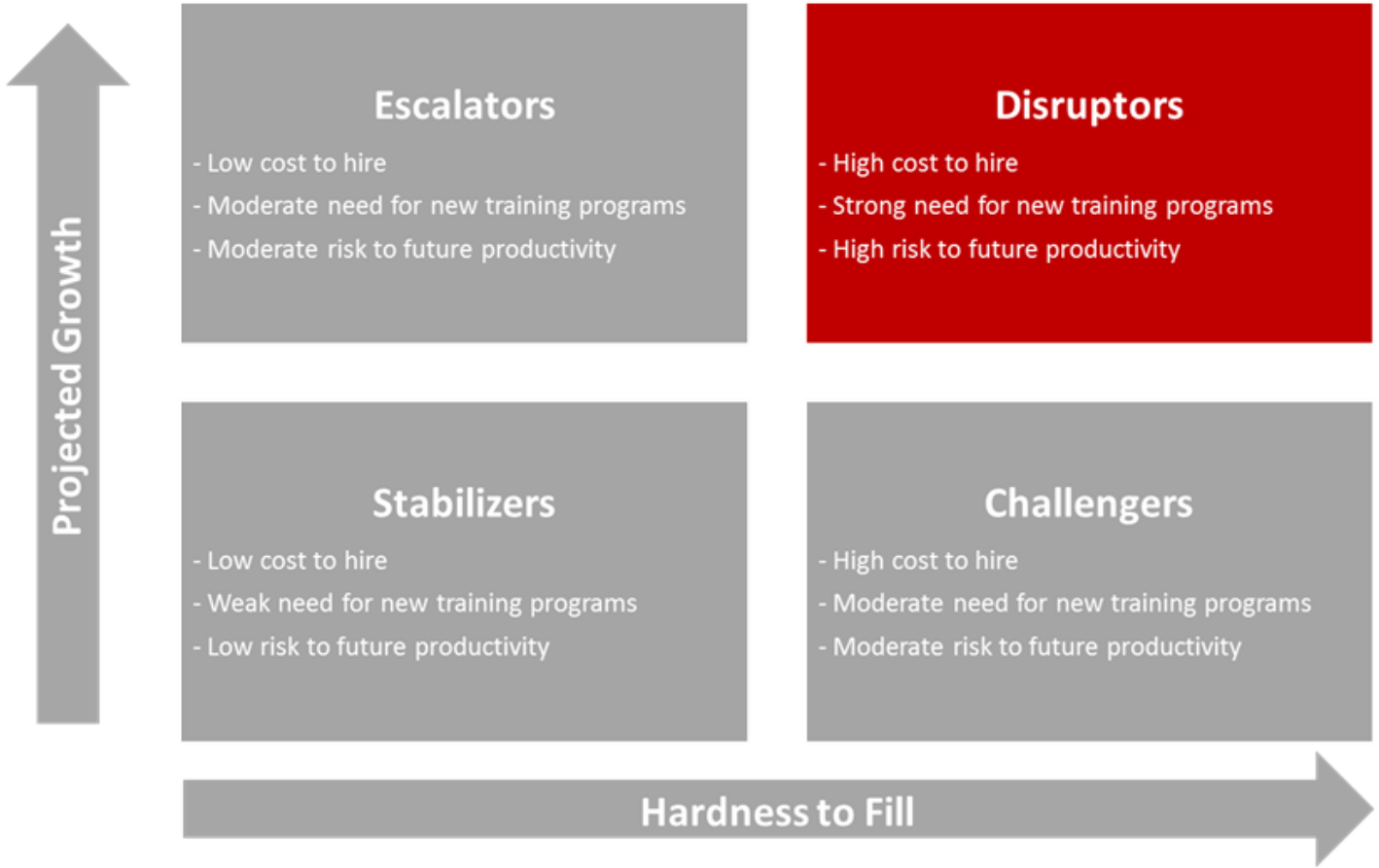


European Training Foundation

Identification et priorisation

des emplois et des compétences perturbateurs

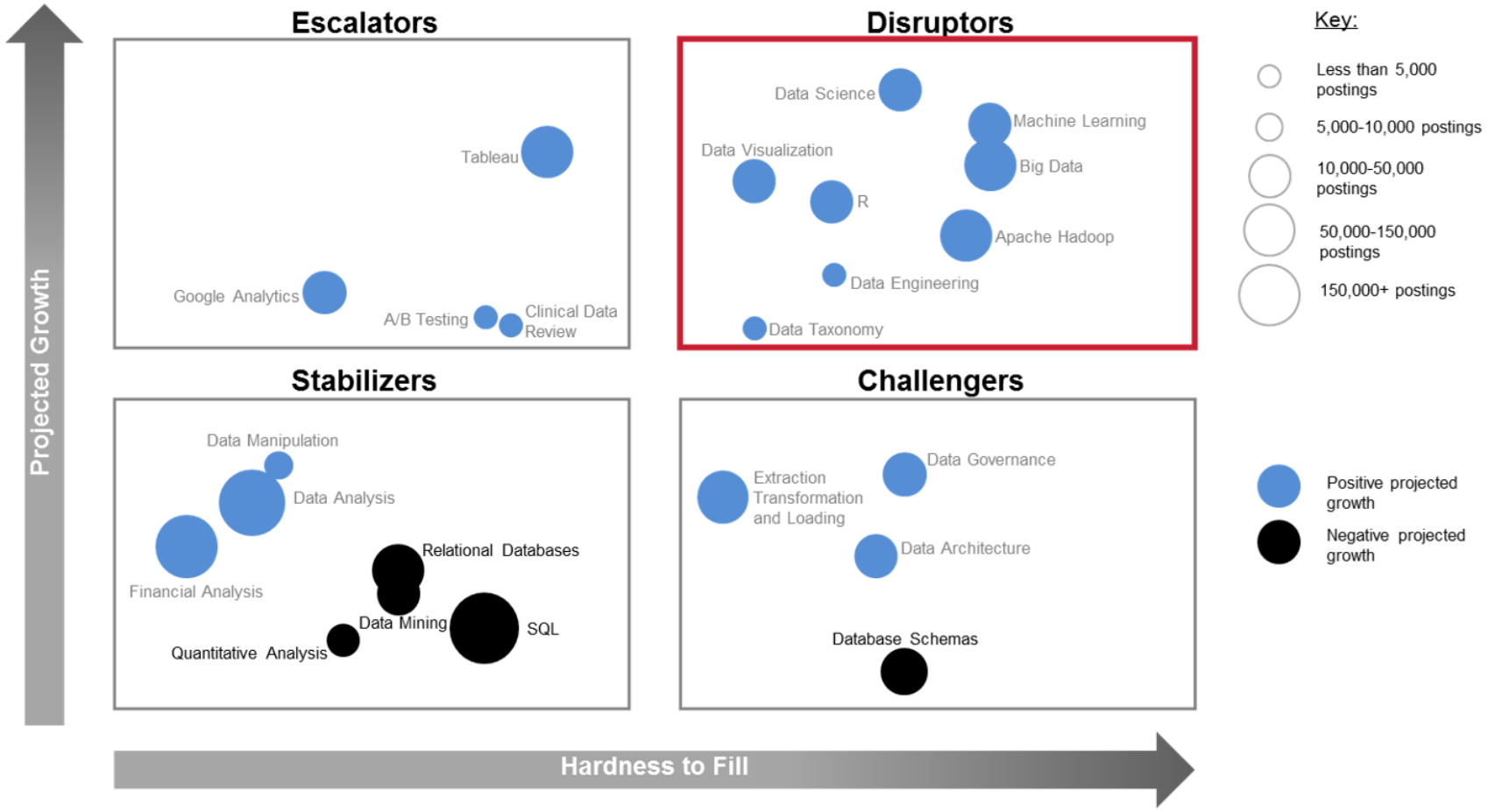
Cadre pour identifier et anticiper les compétences susceptibles de constituer un défi pour le marché à l'avenir.



ETF Working together Learning for life
European Training Foundation

Même dans des domaines émergents comme la science des données

les compétences ne sont pas toutes créées de manière égale.

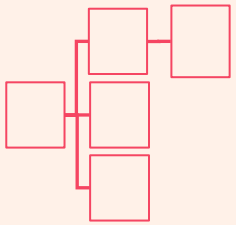


ETF Working together Learning for life
European Training Foundation

Connexion de plusieurs sources

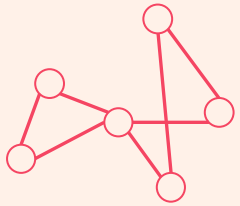
Lightcast

Données structurées vs Big Data



Données structurées

Les données collectées et rassemblées dans un but précis sont présentées dans une structure claire et ordonnée. Dans le cas d'Emsi, il s'agit généralement de données issues d'enquêtes statistiques gouvernementales, conçues pour poser des questions explicites à des échantillons ciblés de publics spécifiques.



Big Data

Données à très grande échelle saisies à partir de certaines transactions plutôt que dans le cadre d'un exercice spécifique de collecte de données. Dans le cas d'Emsi, cela se traduit par la collecte d'offres d'emploi et de profils de travailleurs à partir de différentes sources Web.

Données agrégées vs microdonnées

Données agrégées

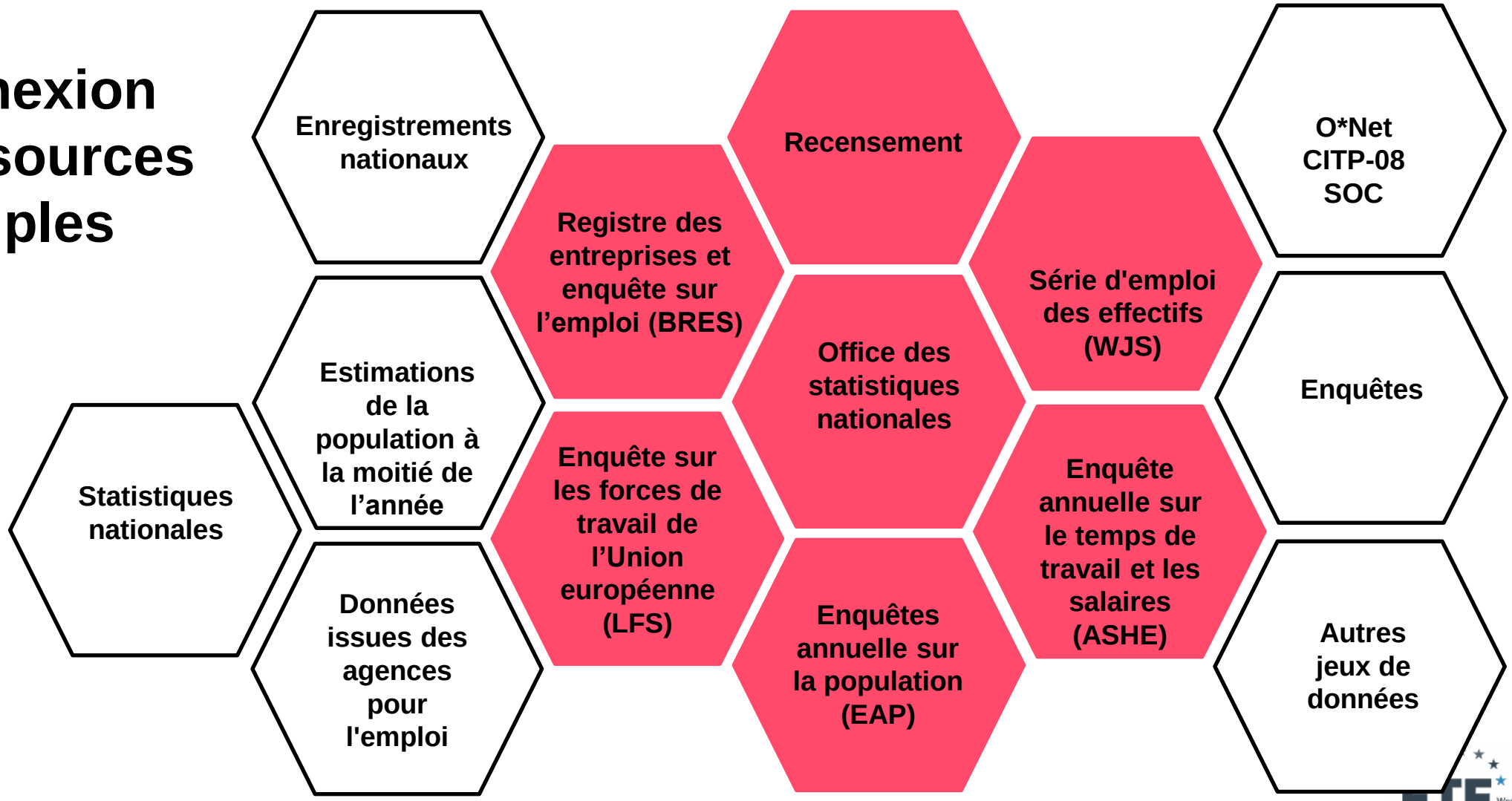
Les données agrégées nous parviennent sous forme de statistiques synthétiques : le nombre total de X, le niveau moyen de Y, le pourcentage de Z. De nombreuses données structurées ne sont disponibles que sous forme agrégée, l'accès aux microdonnées étant limité.

Micro donnée

Les microdonnées ont plus d’impact, les points d’information sur chaque cas individuel étant disponibles sans agrégation ; nous pouvons ensuite les découper comme nous voulons. Les Big Data nous parviennent sous forme de microdonnées, bien que nous les préparions souvent sous des formes standards d'agrégation.

Lightcast

Connexion des sources multiples



ETF Working together Learning for life
European Training Foundation

Lightcast

Analyse des offres d'emploi et des profils - IMT basée sur des documents

- **Les offres d'emploi et les analyses de profils sont des IMT basées sur les documents :** il s'agit d'une source de Big Data où nous maîtrisons l'ensemble du processus, y compris les données brutes.

- **L'analyse des offres d'emploi se fait en temps (presque) réel :** nous récupérons les offres d'emploi tous les jours et dans un délai d'un ou deux jours, nous pouvons utiliser les informations qu'elles produisent.

- **L'analyse des offres d'emploi est très riche :** des millions d'offres d'emploi ont été publiées au cours de ces 2 dernières années (et leur nombre s'accroît jour après jour).

- **L'analyse des offres d'emploi est bruitée :** les offres d'emploi peuvent évoluer différemment en fonction de la demande sous-jacente du marché du travail ; elles doivent donc être traitées en connaissance de cause.

ETF Working together Learning for life
European Training Foundation

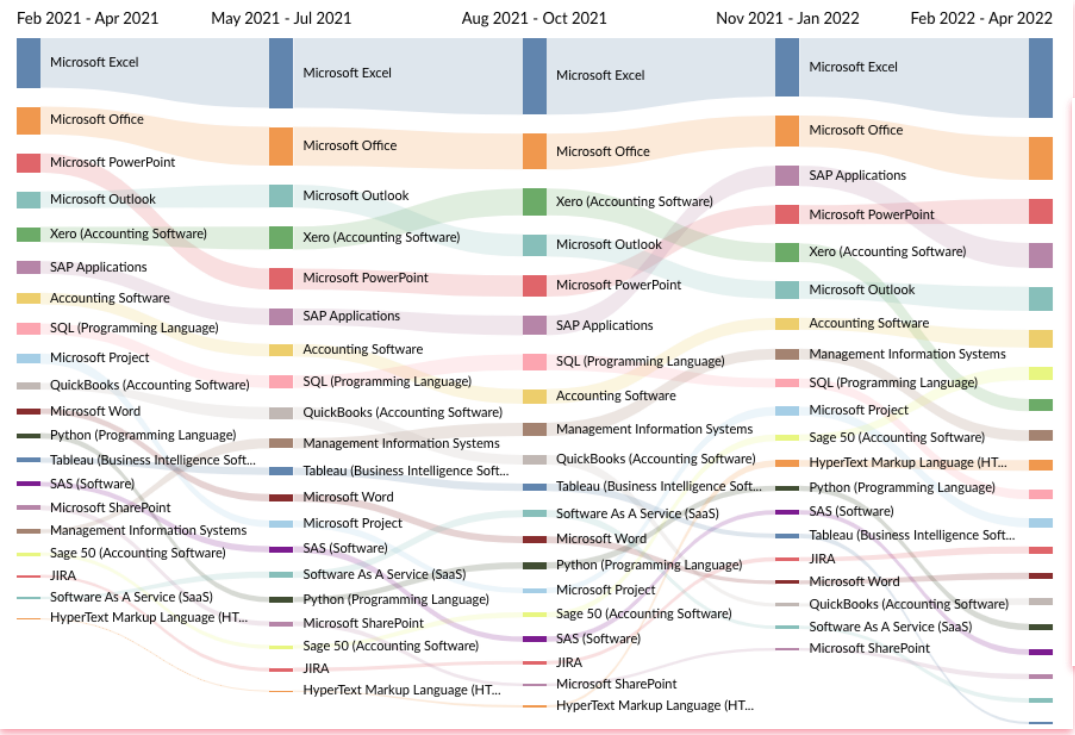
Session virtuelle

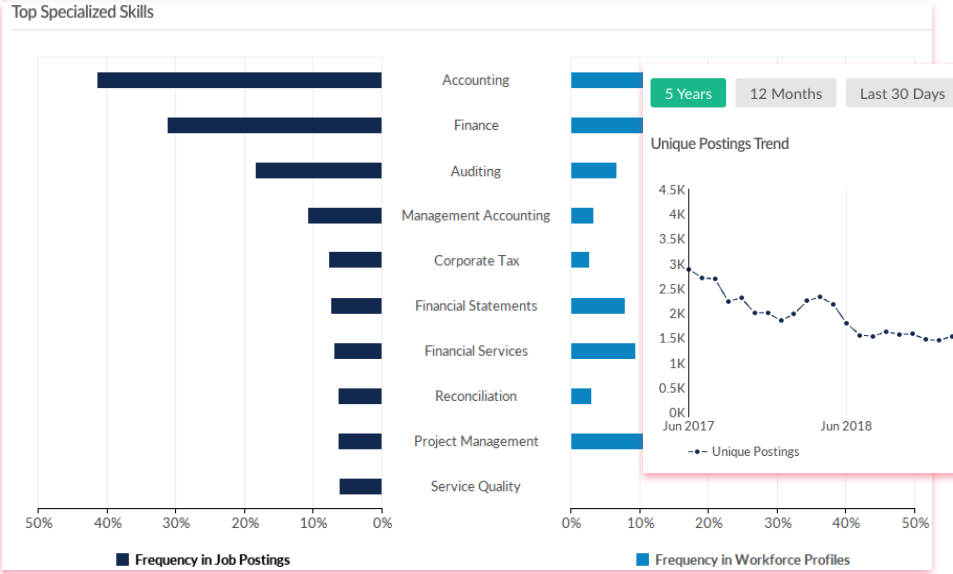
ETF Working together Learning for life
European Training Foundation

Tendances des compétences | Déficit de compétences | Activité de recrutement

de Lightcast Analyst (Royaume-Uni)

Exemple : professions mappées au domaine de formation Comptabilité (JACS N400) dans la ville de Bristol, à Bath et dans le North East Somerset, le North Somerset et le South Gloucestershire, avril 2021 - mai 2022.

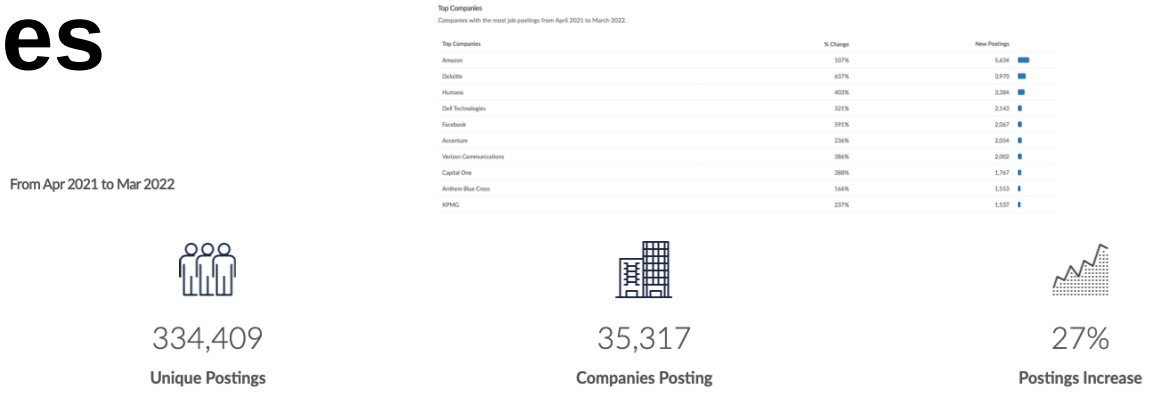




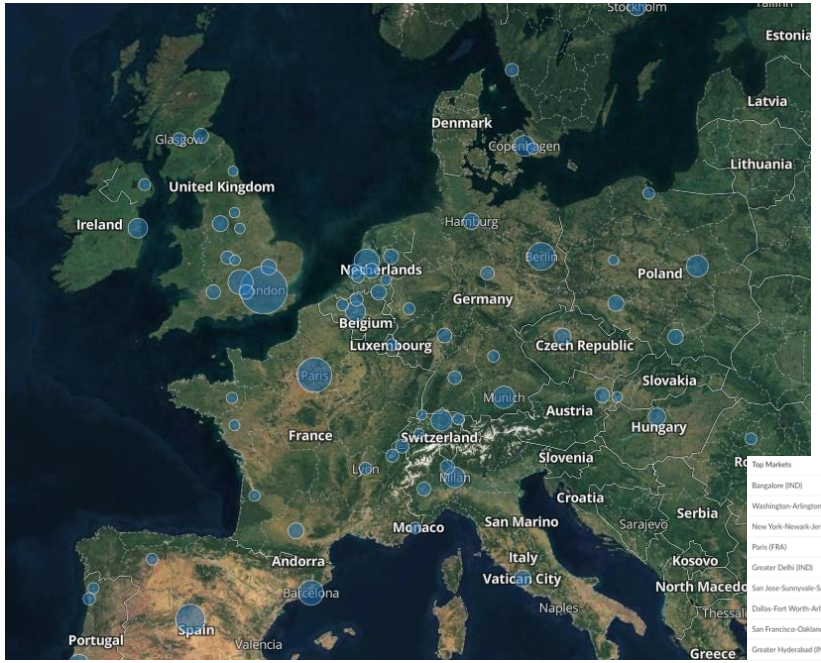
ETF Working together Learning for life
European Training Foundation

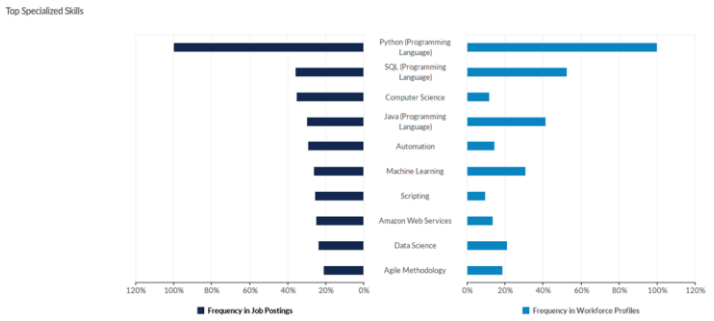
Informations des compétences multinationales de Lightcast Global

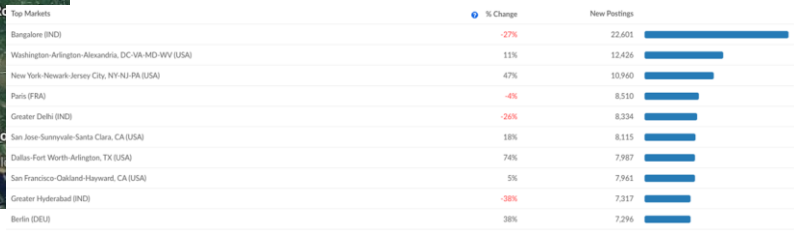
Exemple : analystes des données et experts en mégadonnées ayant des compétences de programmation avec Python



Identifier les concentrations d'offre et de demande de compétences dans les groupes professionnels au niveau mondial. Creuser dans les données pour glaner des informations sur les compétences les plus importantes d'une population en particulier, les principaux employeurs et les estimations de la population.





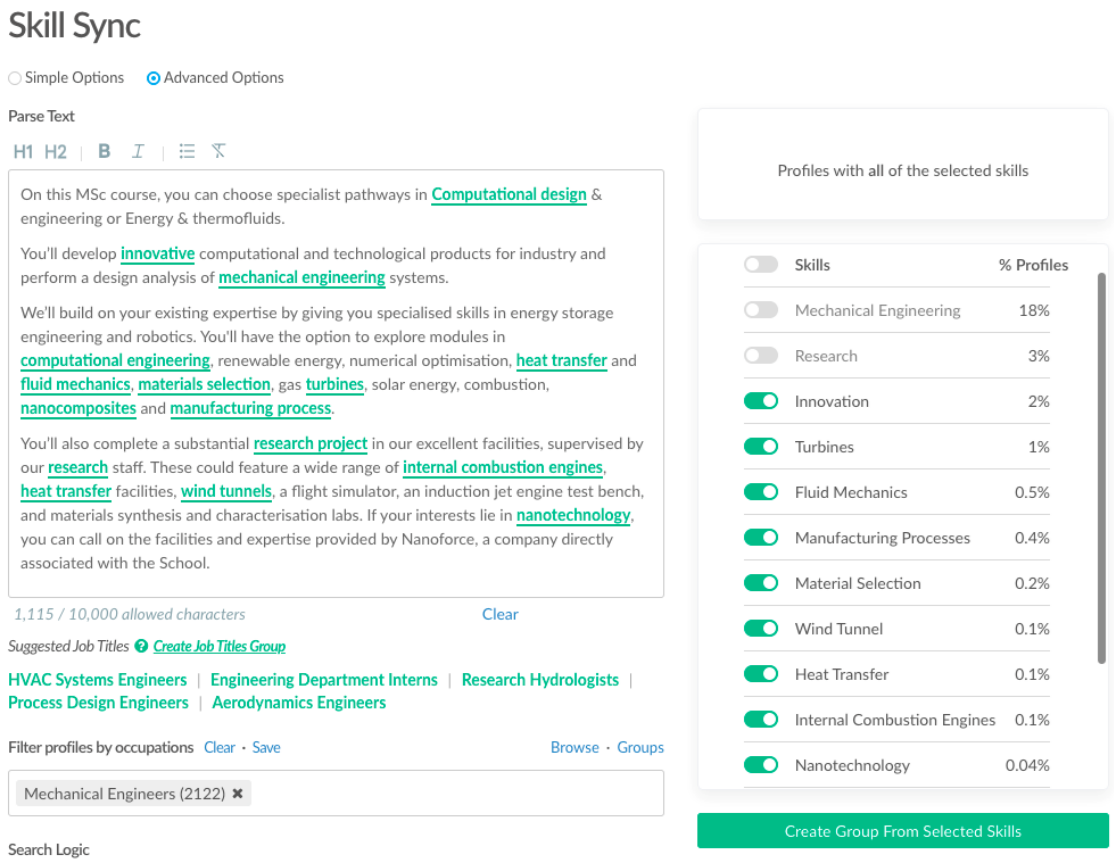


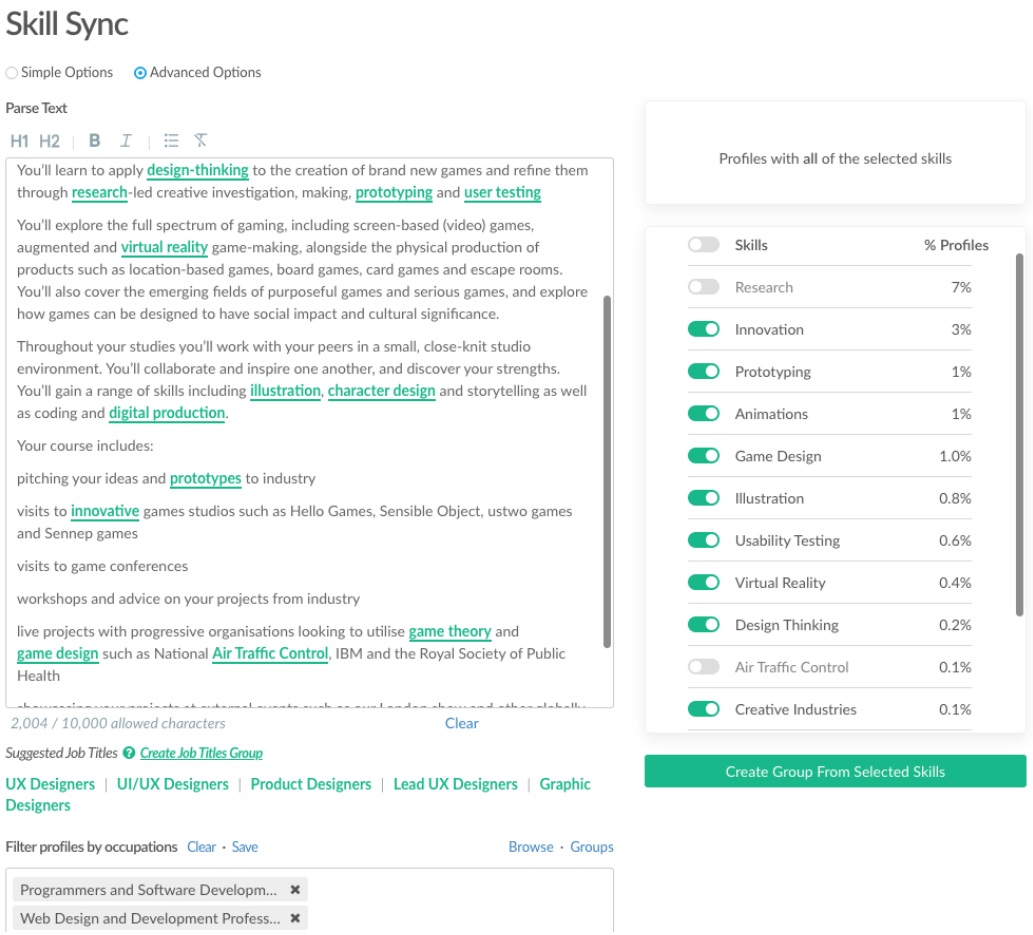
Top Markets	% Change	New Postings
Bangalore (IND)	-27%	22,601
Washington-Arlington-Alexandria, DC-VA-MD-WV (USA)	11%	12,426
New York-Newark-Jersey City, NY-NJ-PA (USA)	47%	10,960
Paris (FRA)	-4%	8,510
Greater Delhi (IND)	-26%	8,334
San Jose-Sunnyvale-Santa Clara, CA (USA)	18%	8,115
Dallas-Fort Worth-Arlington, 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

Examen des données sur les compétences issues des documents de formation

de Lightcast Skill Sync

Ex. : ingénieur mécanicien spécialisé MSc, QMUL et diplôme BA conception de jeux et art,





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...

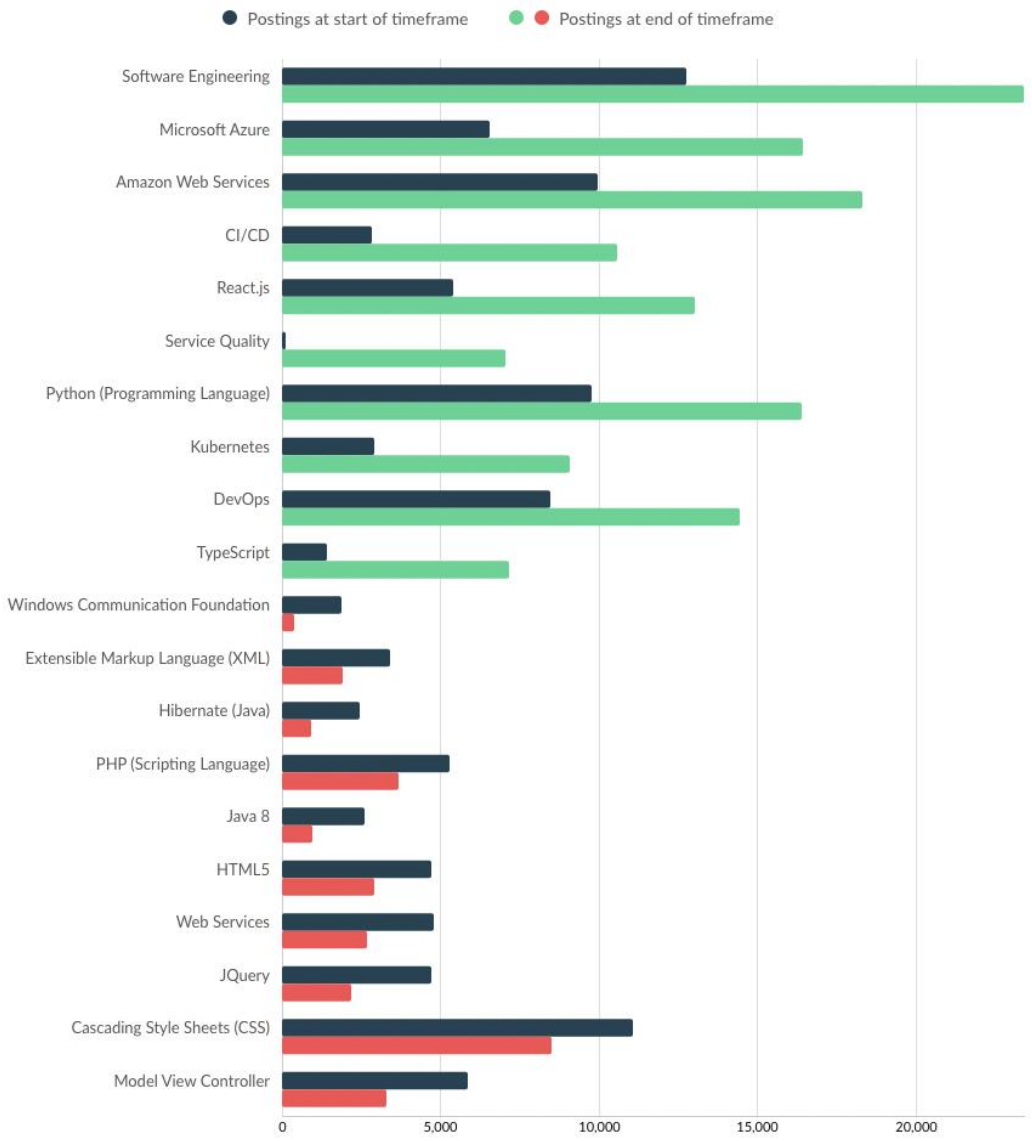
Exclude Skills & Qualifications Groups

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