Big Data for Labour Market Intelligence

Capacity development programme 2022

Module 1: Technical training

Session 1

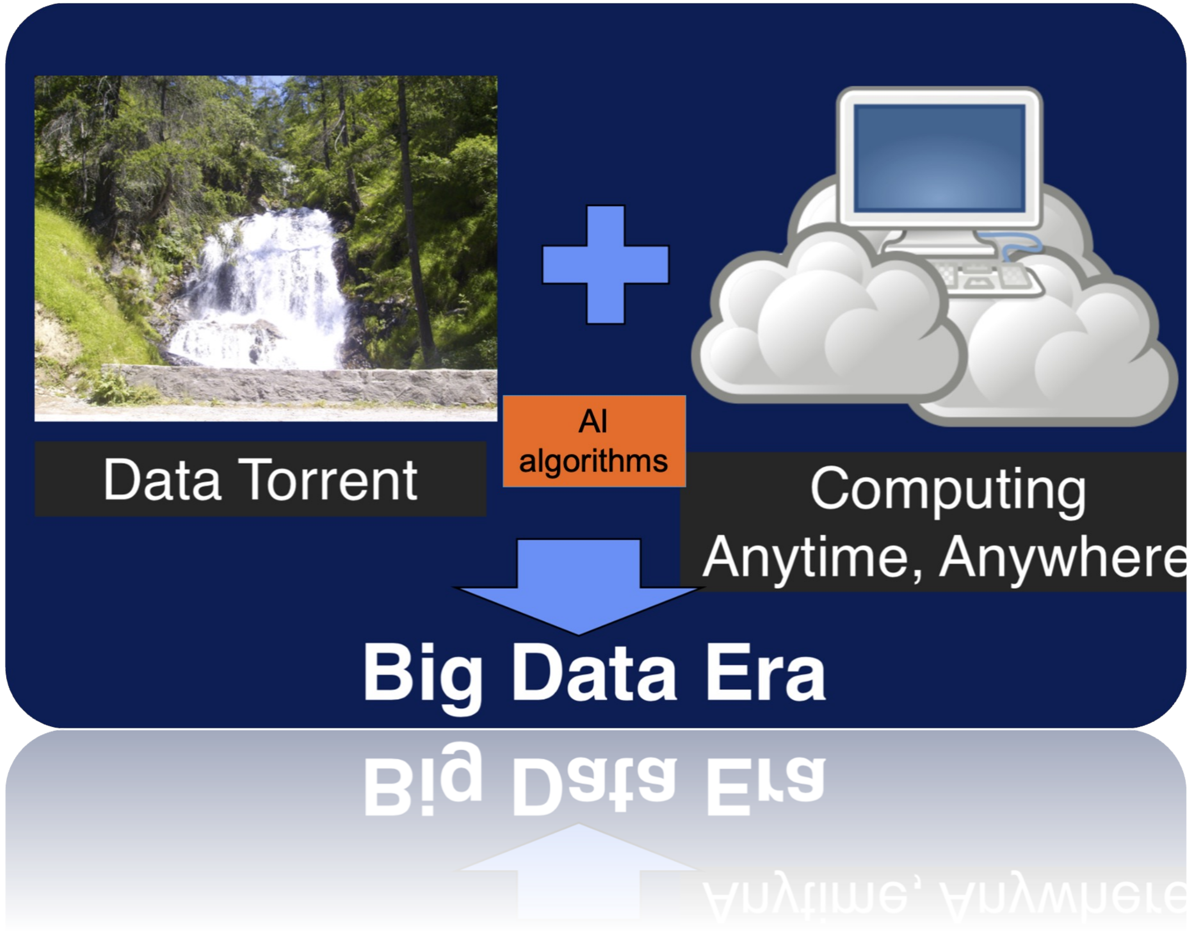
Online Job Vacancy analysis: innovation in LMI – overview

Speaker: Mauro Pelucchi

02/11/2022

ETF project

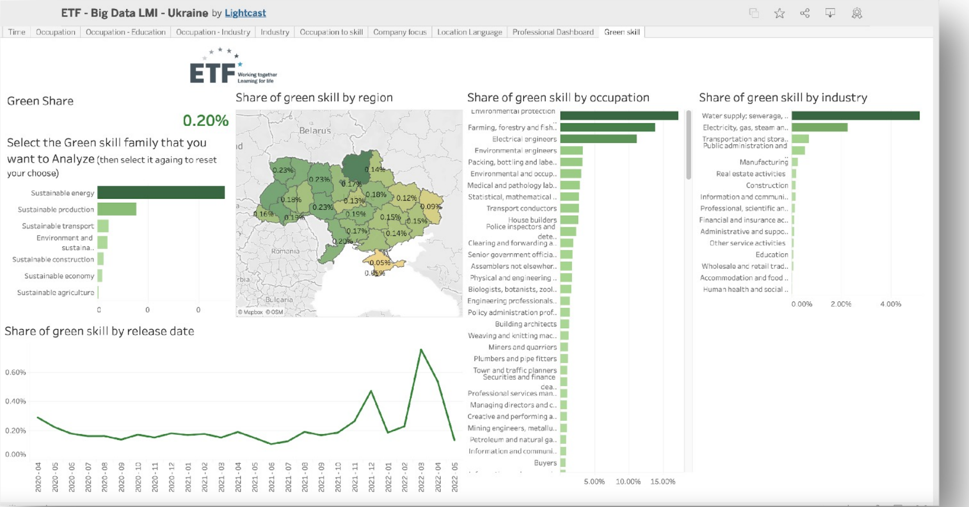
Big Data for LMI 2018-2021



- **2018-2019: Methodology**: first step - brief methodological handbook “Big Data for labour market intelligence: an introductory guide” (published in **2019**).
- **2019: First application**: Feasibility analysis – Landscaping of Web Labour Markets Tunisia and Morocco
- **2019-2021:** 3 main training programmes for experts of the partner countries
- **2020:** Creation of the complete OJV analysis system and dashboards: Tunisia and Ukraine
  - Analytical reports: LM and skills Ukraine and Tunisia
- **2021**:
  - New country – Georgia;
  - **Green dashboard;**
- The data system is based exclusively on **demand** – based on **job vacancies (OJV)** posted on web portals
- Full comparability with the Real-Time data system of the EU-27 (same methodology)
- ETF works with the data analytics specialists of **University Milano-Bicocca and Lightcast**

ETF project Big data for labour market information

2022: new phase



ETF - Big Data LMI - Ukraine by Lightcast

Green Share

0.20%

Select the Green skill family that you want to Analyze (then select it againg to reset your choose)

Time
Occupation
Occupation - Education
Occupation - Industry
Industry
Occupation to skill
Company focus
Location Language
Professional Dashboard
Green skill
ETF Working together Learning for life
Sustainable energy
Sustainable production
Sustainable transport
Environment and sustaina..
Sustainable construction
Sustainable economy
Sustainable agriculture
0
0
0
Share of green skill by region
Belarus
Romania
Bulgaria
0.23%
0.23%
0.14%
0.17%
0.18%
0.13%
0.18%
0.12%
0.09%
0.23%
0.16%
0.19%
0.19%
0.15%
0.15%
0.17%
0.14%
0.20%
0.05%
0.05%
© Mapbox © OSM
Share of green skill by occupation
Environmental protection
Farming, forestry and fish..
Electrical engineers
Environmental engineers
Packing, bottling and labe..
Environmental and occup..
Medical and pathology lab..
Statistical, mathematical ..
Transport conductors
House builders
Police inspectors and dete..
Clearing and forwarding a..
Senior government officia..
Assemblers not elsewher..
Physical and engineering ..
Biologists, botanists, zool..
Engineering professionals..
Policy administration prof..
Building architects
Weaving and knitting mac..
Miners and quarriers
Plumbers and pipe fitters
Town and traffic planners
Securities and finance dea..
Professional services man..
Managing directors and c..
Creative and performing a..
Mining engineers, metallu..
Petroleum and natural ga..
Information and communi..
Buyers
5.00%
10.00%
15.00%
Share of green skill by industry
Water supply; sewerage, ..
Electricity, gas, steam an..
Transportation and stora..
Public administration and ..
Manufacturing
Real estate activities
Construction
Information and communi..
Professional, scientific an..
Financial and insurance ac..
Administrative and suppo..
Other service activities
Education
Wholesale and retail trad..
Accommodation and food ..
Human health and social ..
0.00%
2.00%
4.00%
Share of green skill by release date
0.60%
0.40%
0.20%
0.00%
2020 - 04
2020 - 05
2020 - 06
2020 - 07
2020 - 08
2020 - 09
2020 - 10
2020 - 11
2020 - 12
2021 - 01
2021 - 02
2021 - 03
2021 - 04
2021 - 05
2021 - 06
2021 - 07
2021 - 08
2021 - 09
2021 - 10
2021 - 11
2021 - 12
2022 - 01
2022 - 02
2022 - 03
2022 - 04
2022 - 05

- **Continue, maintain, improve – the 3 existing country systems**
- **Expand to new countries**
- **Ukraine:**
  - ESCO – translation to Ukrainian language and launch on ESCO Platform; mapping to national classifications (occupations, skills)
  - Lviv project – focused on 1 region
  - PES OJV database and dashboard: significant improvements
- **General**:
  - Concept paper: contextualise OJV data in the wider LMI data – because OJV adds value to other reliable data sources
  - Capacity development, dissemination of results

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)

→ Estratégia de Comunicação para toda a gente conhecer o QNQ e o CNQ?
① QNQ e o CNQ devem estar ligados ao Sistema de Certificação, aos utilizadores, empresas, etc

Topics

- What is Labour Market Intelligence?
- New sources, why?
- Big data for LMI
- Methodology

Q: Do You Know the Emerging Skills In Your Labor Market?

Q: Do You Know Your Local *Skill* Gaps and What To Do About Them?

Continuously evolving Labour Market

Context

Digitalization of professions

Relevance of Soft skills

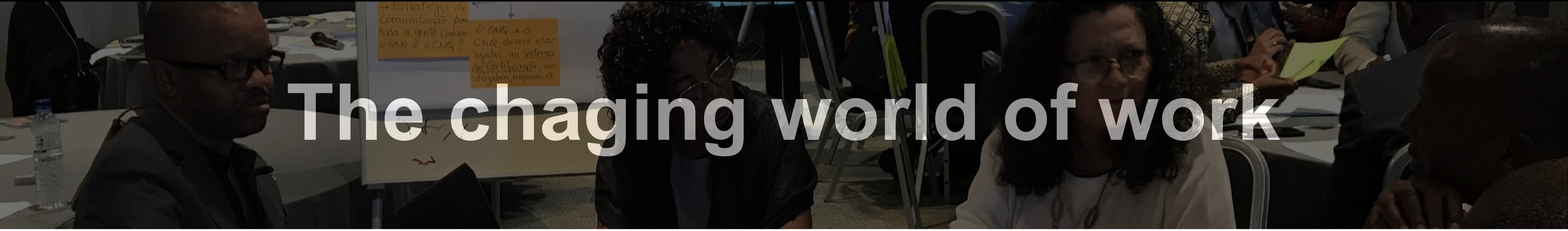
Internationalisation

New professions and skills emerging

Smart and Remote working

Impact of Covid-19 pandemic

Green transition



The chaging world of work

A shared language between employers and job seekers:

- Employers post job openings with increasingly specific skill requirements to attract talent they need
- Job seekers create online profiles and resumes with increasingly skill descriptions to market themselves to potential employers

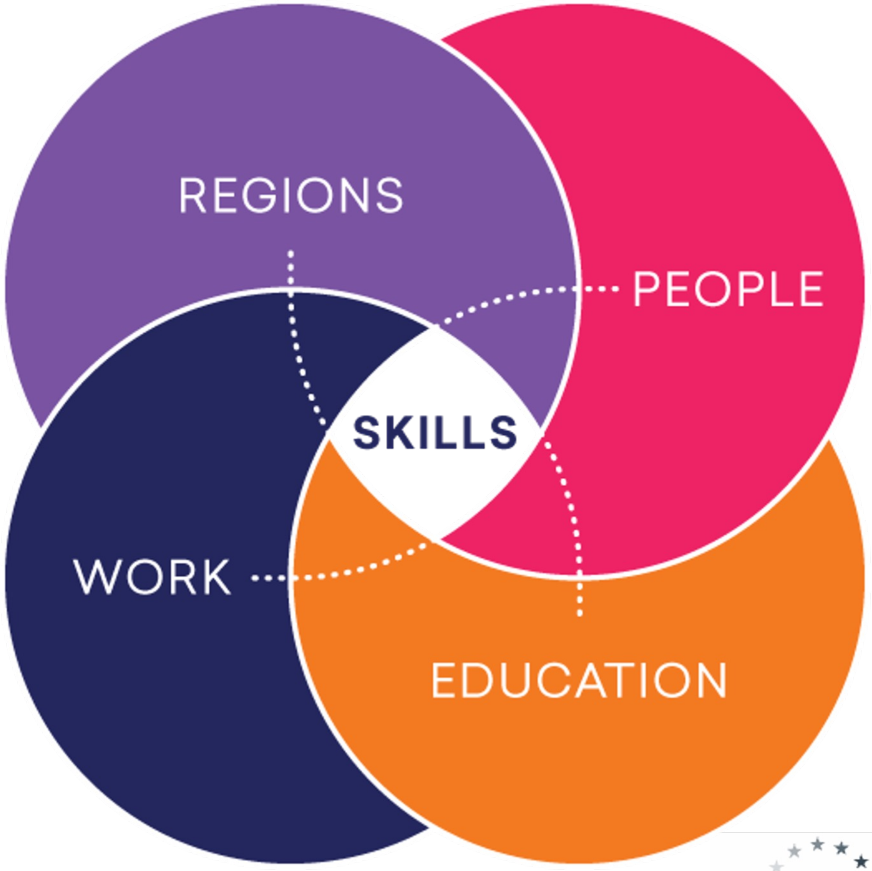
We're in a skill-based economy

What is a skill?

- Anything that defines or describes someone’s knowledge and experience
  - Hard Skills
  - Soft (or Essential) Skills
  - Certifications

Why Skills?

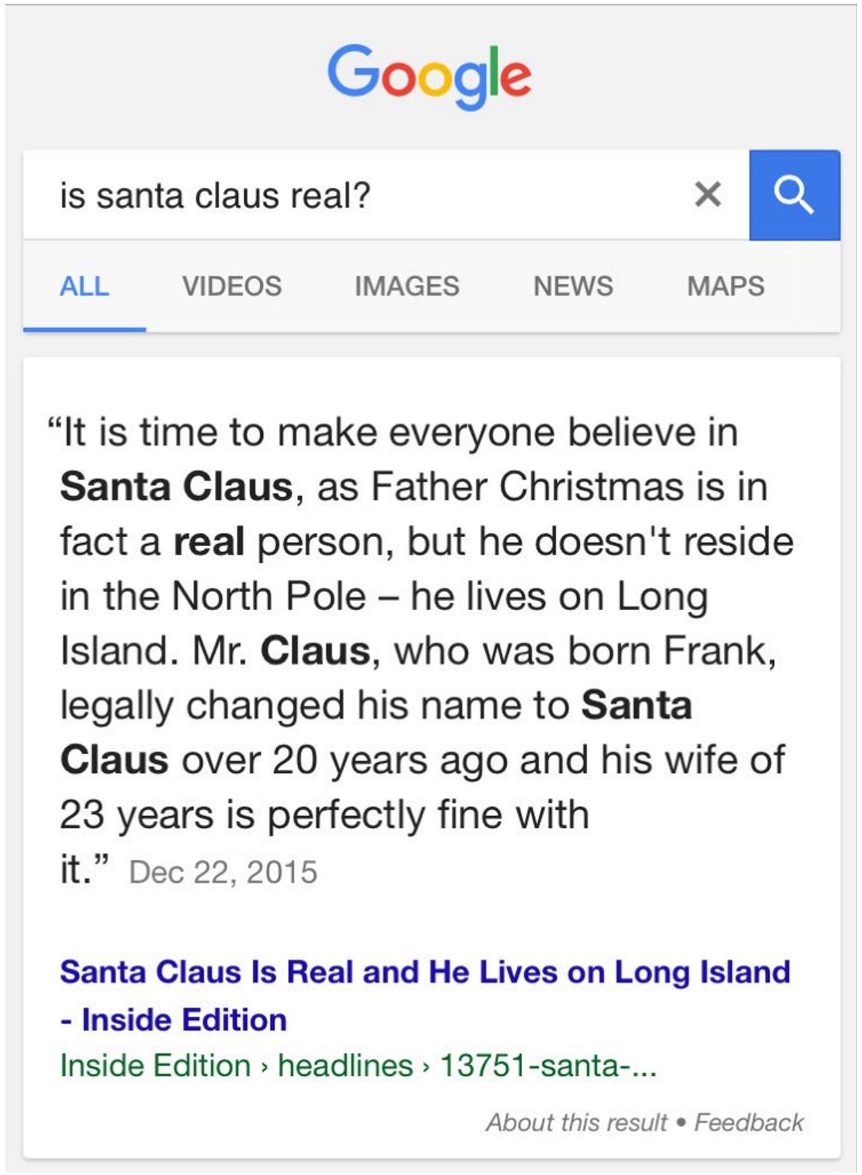
- Common language
- Equity
- Agile and precise
- Better understand talent supply & demand regionally
- Market & match talent to companies



ETF Working together Learning for life

European Training Foundation

New questions



Google

is santa claus real?

ALL VIDEOS IMAGES NEWS MAPS

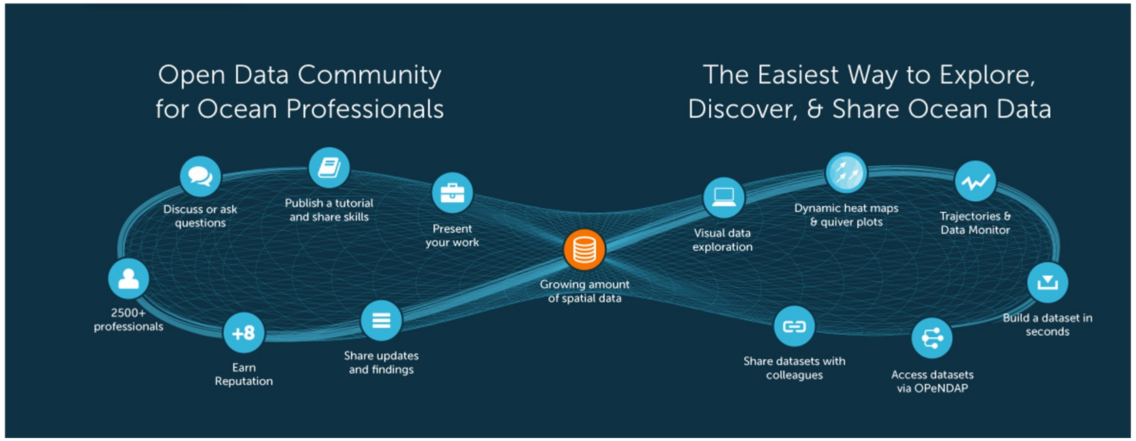
"It is time to make everyone believe in **Santa Claus**, as Father Christmas is in fact a **real** person, but he doesn't reside in the North Pole – he lives on Long Island. Mr. **Claus**, who was born Frank, legally changed his name to **Santa Claus** over 20 years ago and his wife of 23 years is perfectly fine with it." Dec 22, 2015

Santa Claus Is Real and He Lives on Long Island - Inside Edition

Inside Edition › headlines › 13751-santa-...

About this result • Feedback

New sources



Open Data Community for Ocean Professionals

The Easiest Way to Explore, Discover, & Share Ocean Data

2500+ professionals

Discuss or ask questions

Publish a tutorial and share skills

Present your work

Earn Reputation

Share updates and findings

Growing amount of spatial data

Visual data exploration

Dynamic heat maps & quiver plots

Trajectories & Data Monitor

Build a dataset in seconds

Share datasets with colleagues

Access datasets via OPeNDAP

airbnb

amazon dash BUTTON

TESLA

flightradar24 LIVE AIR TRAFFIC

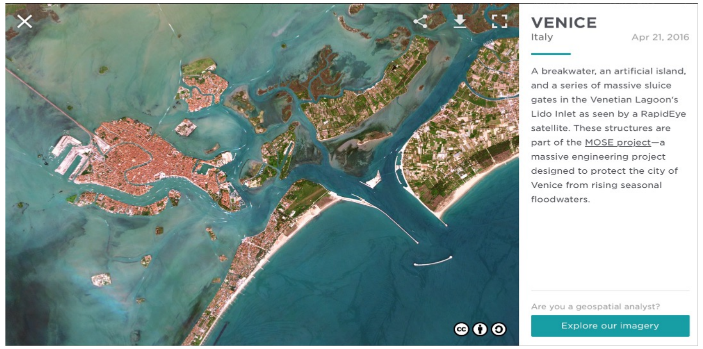
planet.

PLANET OS

Propeller

indeed

Terra Bella



VENICE

Italy Apr 21, 2016

A breakwater, an artificial island, and a series of massive sluice gates in the Venetian Lagoon's Lido Inlet as seen by a RapidEye satellite. These structures are part of the MOSE project—a massive engineering project designed to protect the city of Venice from rising seasonal floodwaters.

Are you a geospatial analyst?

Explore our imagery

myfitnesspal Calorie Counter & Diet Tracker

This is where labour market data is critical!

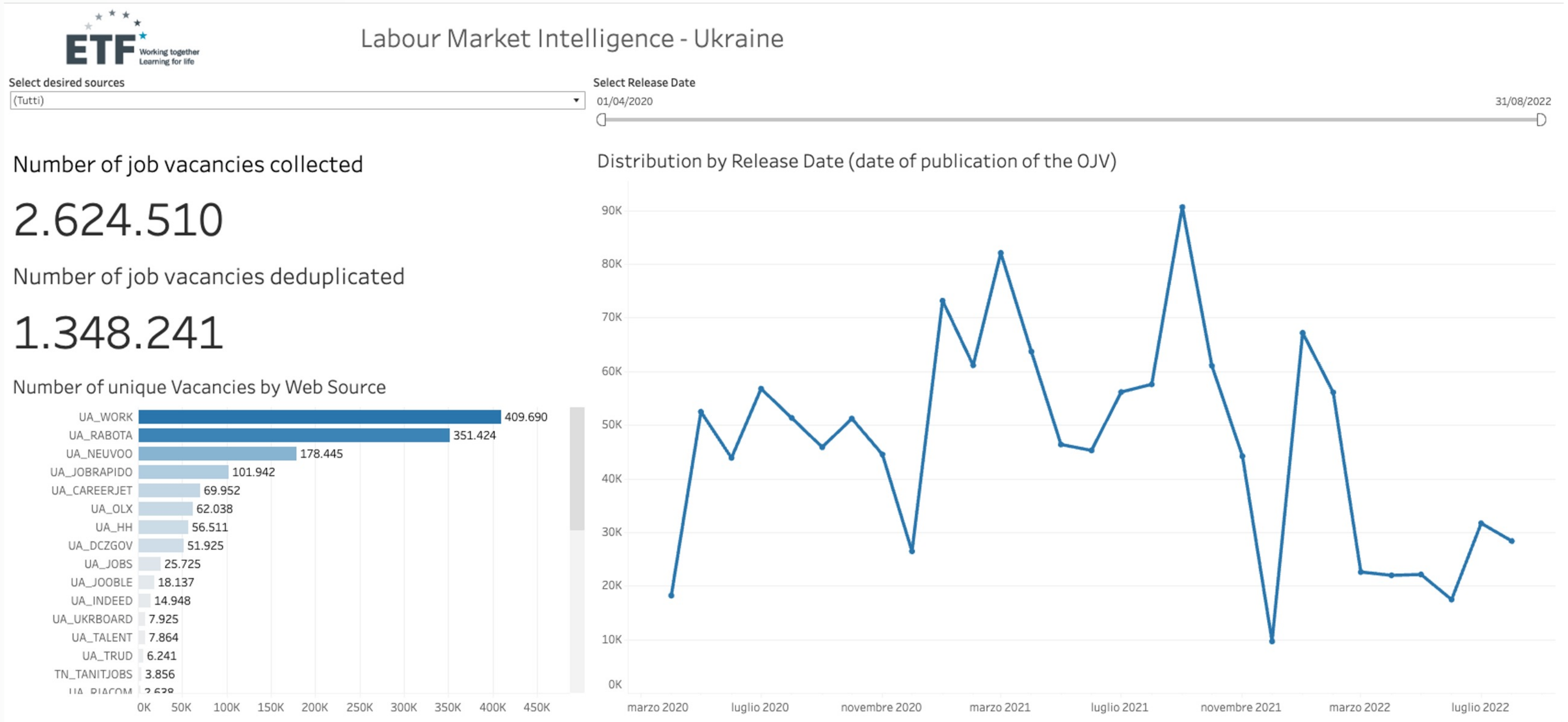
- Official statistics are representative and robust, but can lack detail and timeliness
- They don’t give us the detailed picture, we need:
  - More frequently updated - to track what’s happening now (e.g. Covid-19 Impact analysis)
  - More granular and adherent to real and current market terms - capture emerging trends analyzing what companies are actually looking for

The solution?

Using data derived from online job postings

Real-Time Labour Market Information System on Skill Requirements

Continuously evolving Labour Market



ETF Working together Learning for life

Labour Market Intelligence - Ukraine

Select desired sources: (Tutti)

Select Release Date: 01/04/2020 – 31/08/2022

Number of job vacancies collected

2.624.510

Number of job vacancies deduplicated

1.348.241

Number of unique Vacancies by Web Source

Distribution by Release Date (date of publication of the OJV)

Why Job Posting Labour Market data?

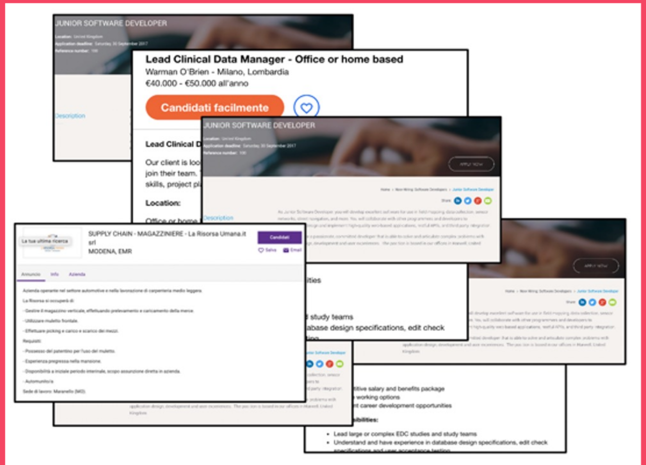
It’s the exact representation of what companies are looking in a given period:

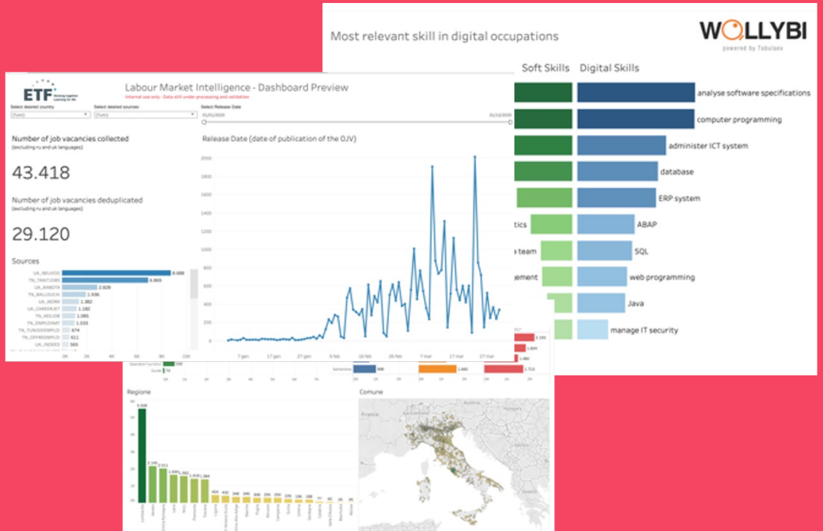
Up to date: companies publish an announcement when they actually need to hire

Detailed: an announcement describes as well as possible the specific need, in terms of:

- Occupation needed
- Requirements (skills, experience, educational level,…)
- Working context (place, contract, sector, working hours,…)

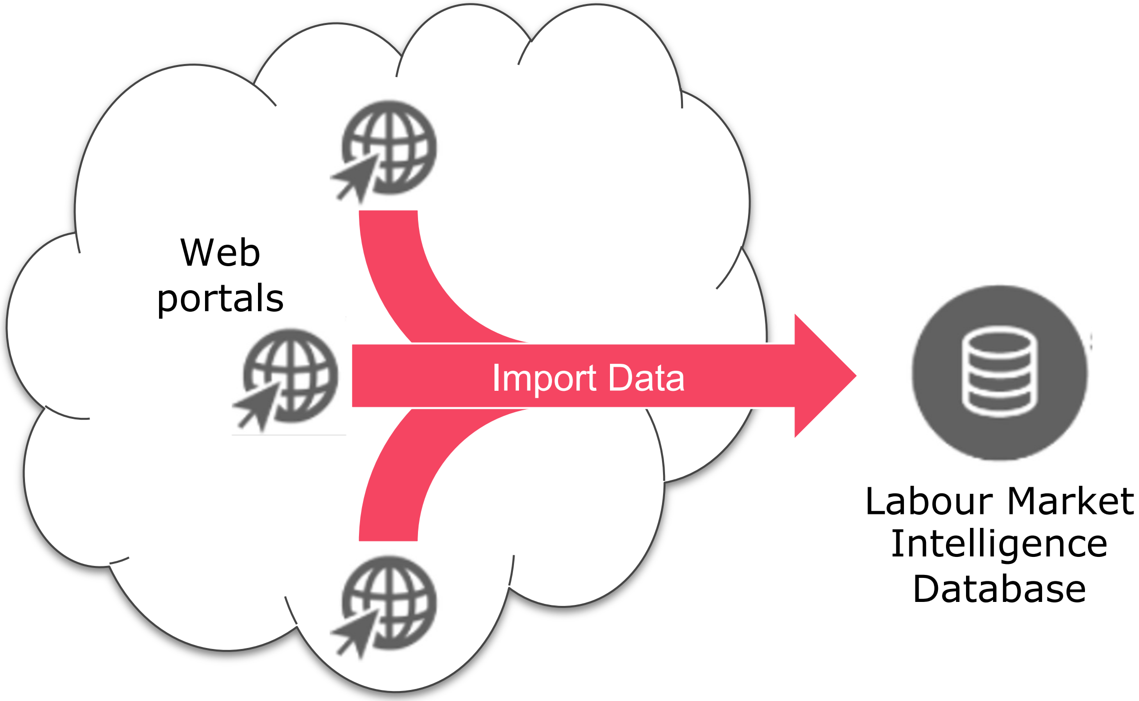
Adherent to reality: market terms are used, both for occupation and skills. This helps identify emerging terminology adopted by Market


JUNIOR SOFTWARE DEVELOPER
Lead Clinical Data Manager - Office or home based
Warman O'Brien - Milano, Lombardia
€40.000 - €50.000 all'anno
Candidati facilmente
JUNIOR SOFTWARE DEVELOPER
SUPPLY CHAIN - MAGAZZINIERE - La Risorsa Umana.it srl
MODENA, EMR


Most relevant skill in digital occupations
WOLLYBI
Soft Skills
Digital Skills
analyse software specifications
computer programming
administer ICT system
database
ERP system
ABAP
SQL
web programming
Java
manage IT security
ETF
Labour Market Intelligence - Dashboard Preview
Number of job vacancies collected
43.418
Number of job vacancies deduplicated
29.120
Sources
Release Date (date of publication of the OJV)
Regione
Comune

New source of data

Web Data ingestion is the process of obtaining and importing data from web portals and storing in a database



What is LMI

Labour Market Intelligence (LMI) is simply insight, information and intelligence about labour markets.

Information on:

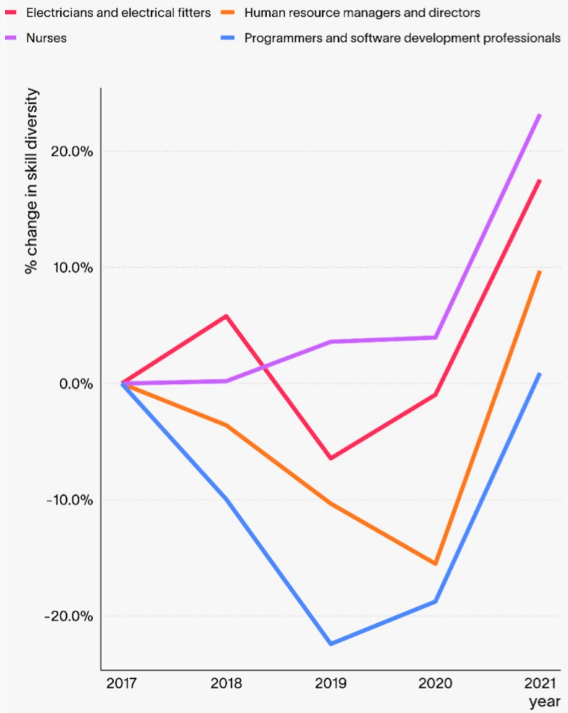
- occupations
- industries
- educational levels for occupations
- workforce demographics

Giving your organisation the peace of mind that its decisions are being made on a basis of solid evidence, rather than assumptions or guesswork

New dimensions and new metrics

Tracking roles' skill range

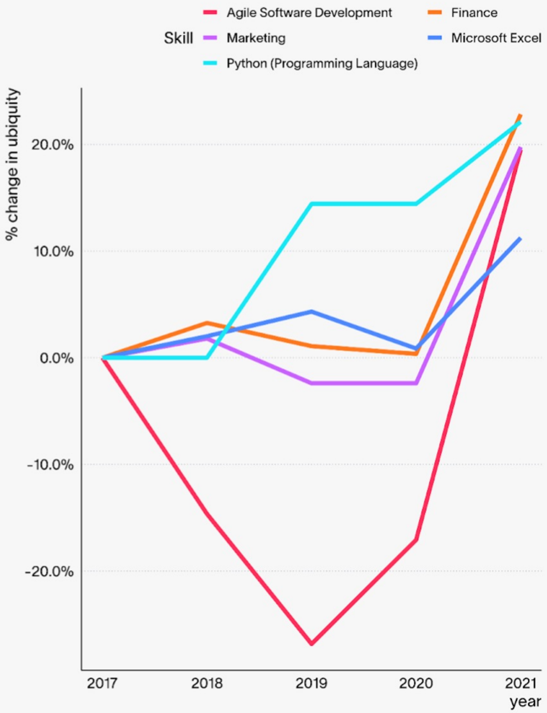
Skill diversity by occupation



Source: Lightcast Job Posting Analytics

Tracking skill ubiquity over time

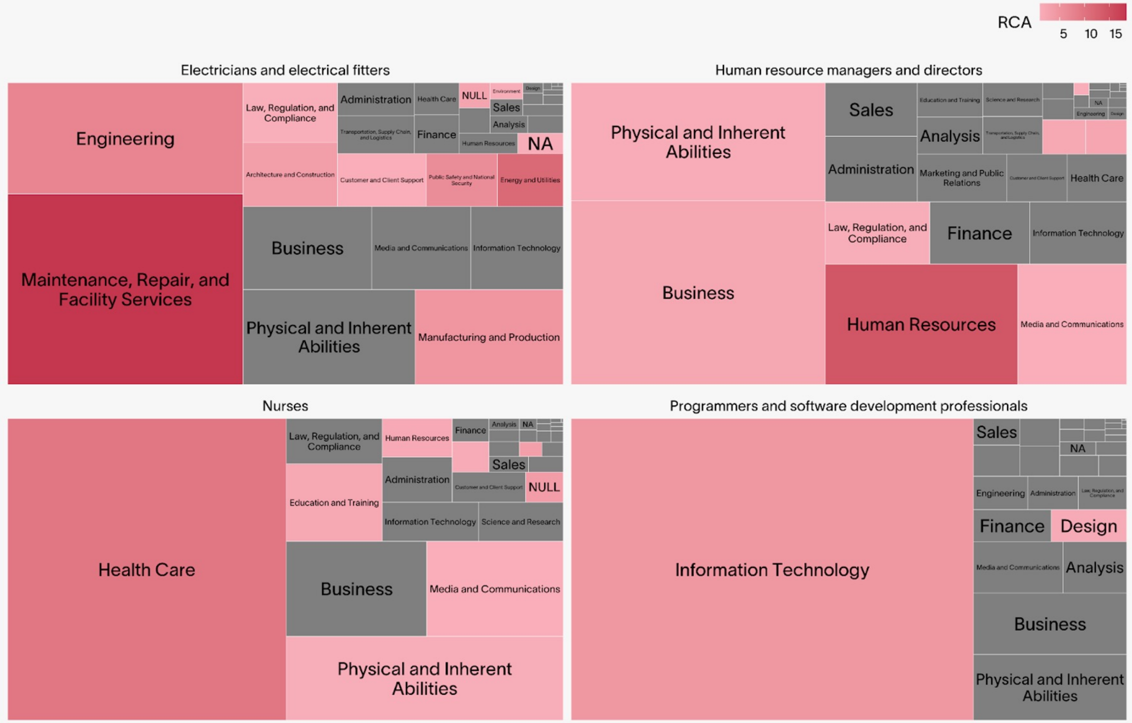
Skill ubiquity across occupations



Source: Lightcast Job Posting Analytics

Describing content using skill category surface and Revealed Comparative Advantage

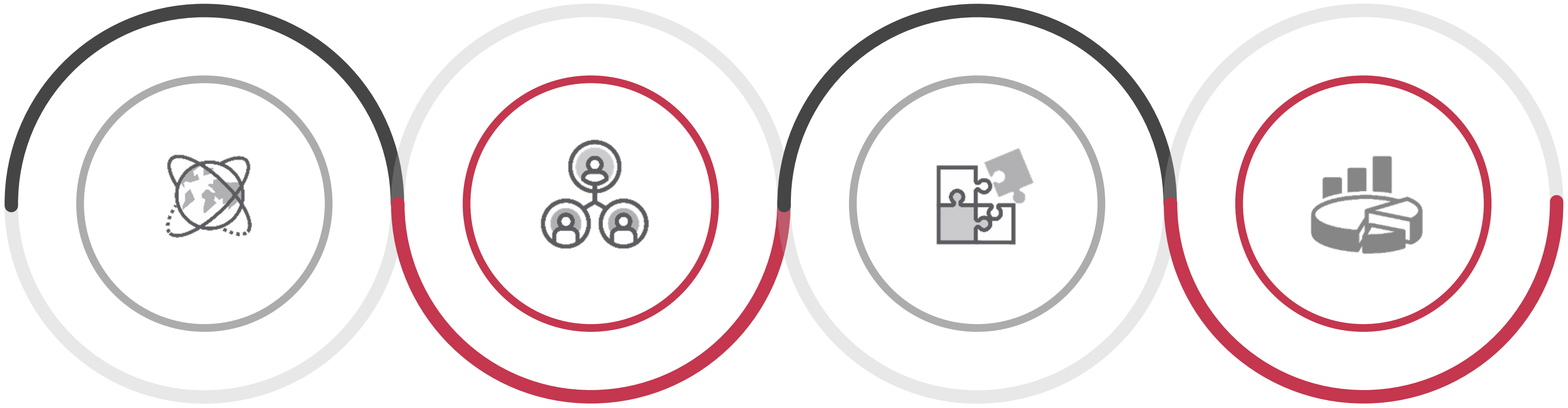
Using skill categories to describe the occupation content (grey is RCA<1)



Source: Lightcast Job Posting Analytics

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

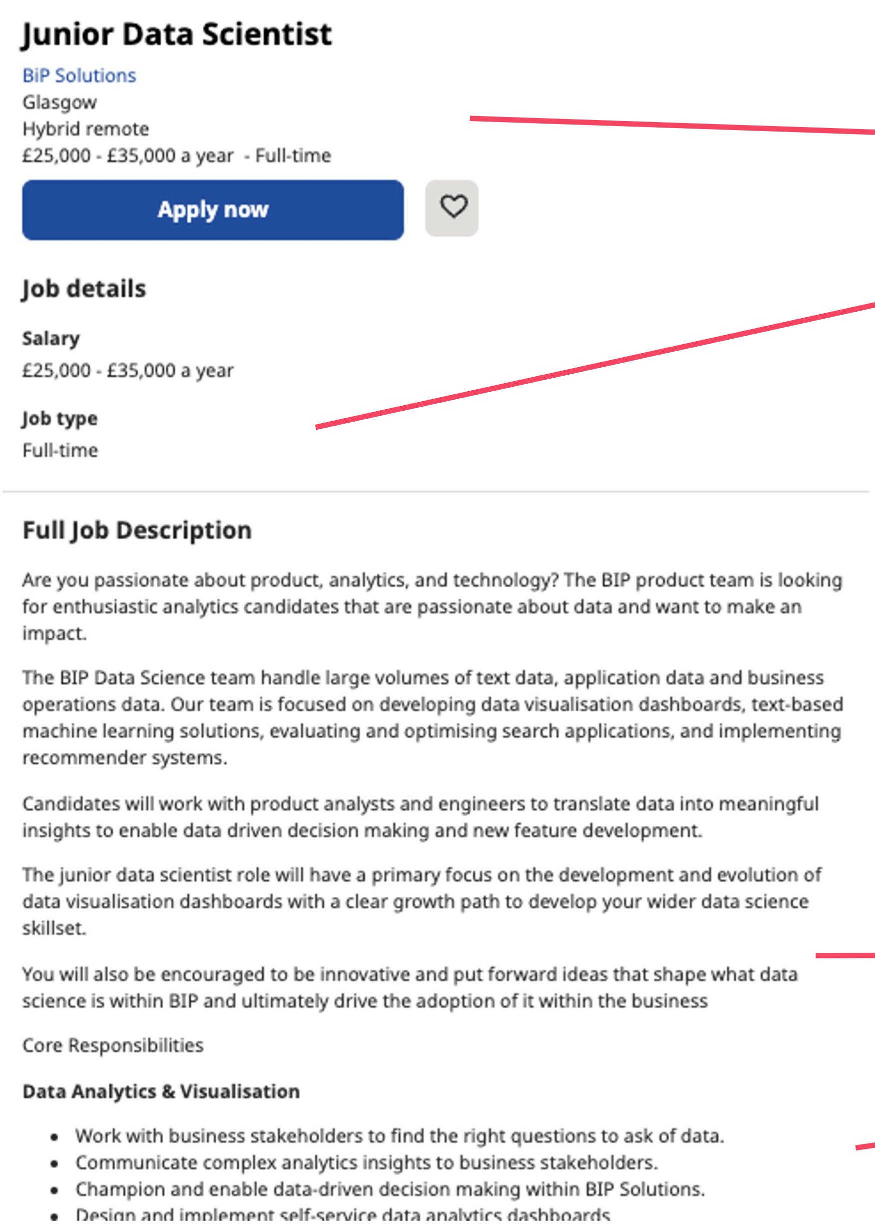
Tagging and structuring

A common language

Data ontology allows for comparisons

Drawing conclusions

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



Junior Data Scientist

BiP Solutions
Glasgow
Hybrid remote
£25,000 - £35,000 a year - Full-time

Apply now

Job details

Salary

£25,000 - £35,000 a year

Job type

Full-time

Full Job Description

Are you passionate about product, analytics, and technology? The BIP product team is looking for enthusiastic analytics candidates that are passionate about data and want to make an impact.

The BIP Data Science team handle large volumes of text data, application data and business operations data. Our team is focused on developing data visualisation dashboards, text-based machine learning solutions, evaluating and optimising search applications, and implementing recommender systems.

Candidates will work with product analysts and engineers to translate data into meaningful insights to enable data driven decision making and new feature development.

The junior data scientist role will have a primary focus on the development and evolution of data visualisation dashboards with a clear growth path to develop your wider data science skillset.

You will also be encouraged to be innovative and put forward ideas that shape what data science is within BIP and ultimately drive the adoption of it within the business

Core Responsibilities

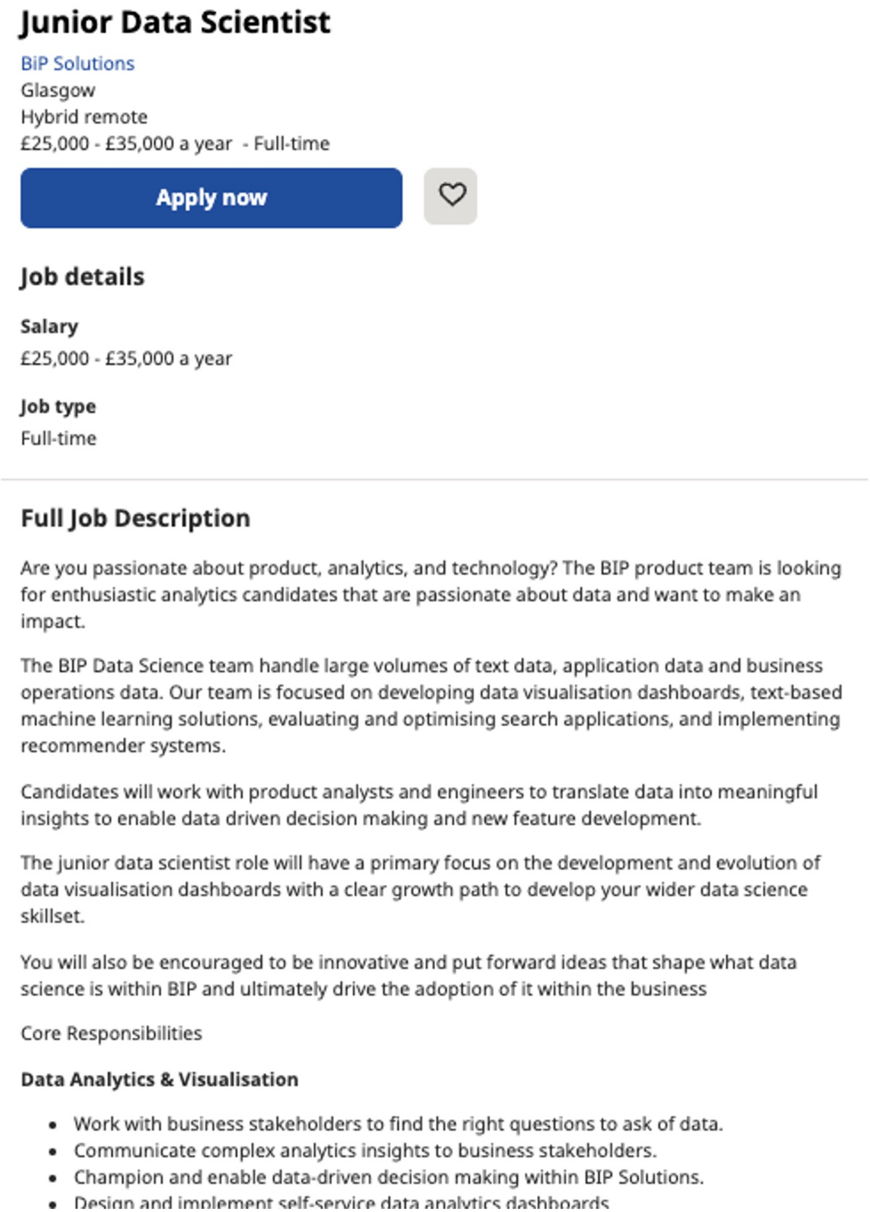
Data Analytics & Visualisation

- Work with business stakeholders to find the right questions to ask of data.
- Communicate complex analytics insights to business stakeholders.
- Champion and enable data-driven decision making within BIP Solutions.
- Design and implement self-service data analytics dashboards

Structured data

Raw data

How do you (a human) classify a job posting in an occupation?



Junior Data Scientist

BiP Solutions
Glasgow
Hybrid remote
£25,000 - £35,000 a year - Full-time

Job details

Salary

£25,000 - £35,000 a year

Job type

Full-time

Full Job Description

Are you passionate about product, analytics, and technology? The BIP product team is looking for enthusiastic analytics candidates that are passionate about data and want to make an impact.

The BIP Data Science team handle large volumes of text data, application data and business operations data. Our team is focused on developing data visualisation dashboards, text-based machine learning solutions, evaluating and optimising search applications, and implementing recommender systems.

Candidates will work with product analysts and engineers to translate data into meaningful insights to enable data driven decision making and new feature development.

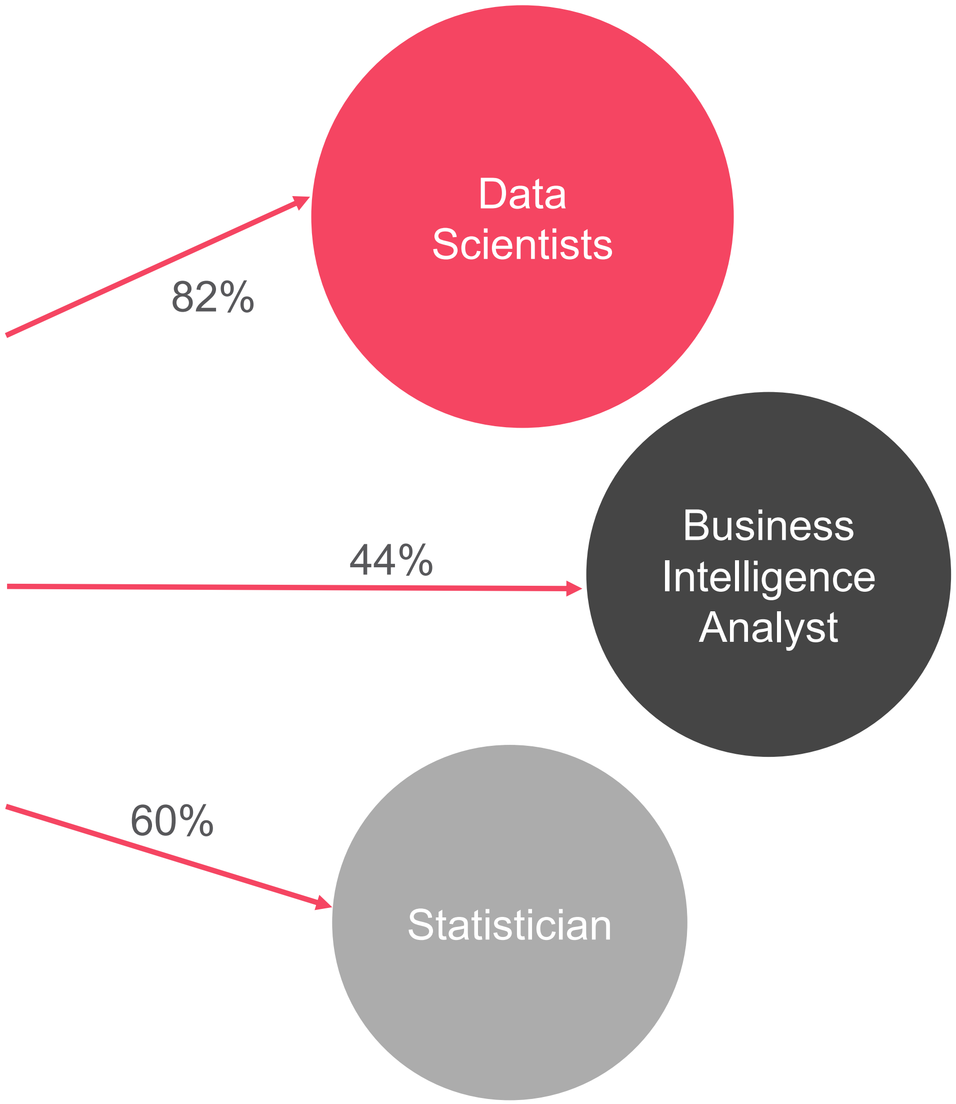
The junior data scientist role will have a primary focus on the development and evolution of data visualisation dashboards with a clear growth path to develop your wider data science skillset.

You will also be encouraged to be innovative and put forward ideas that shape what data science is within BIP and ultimately drive the adoption of it within the business

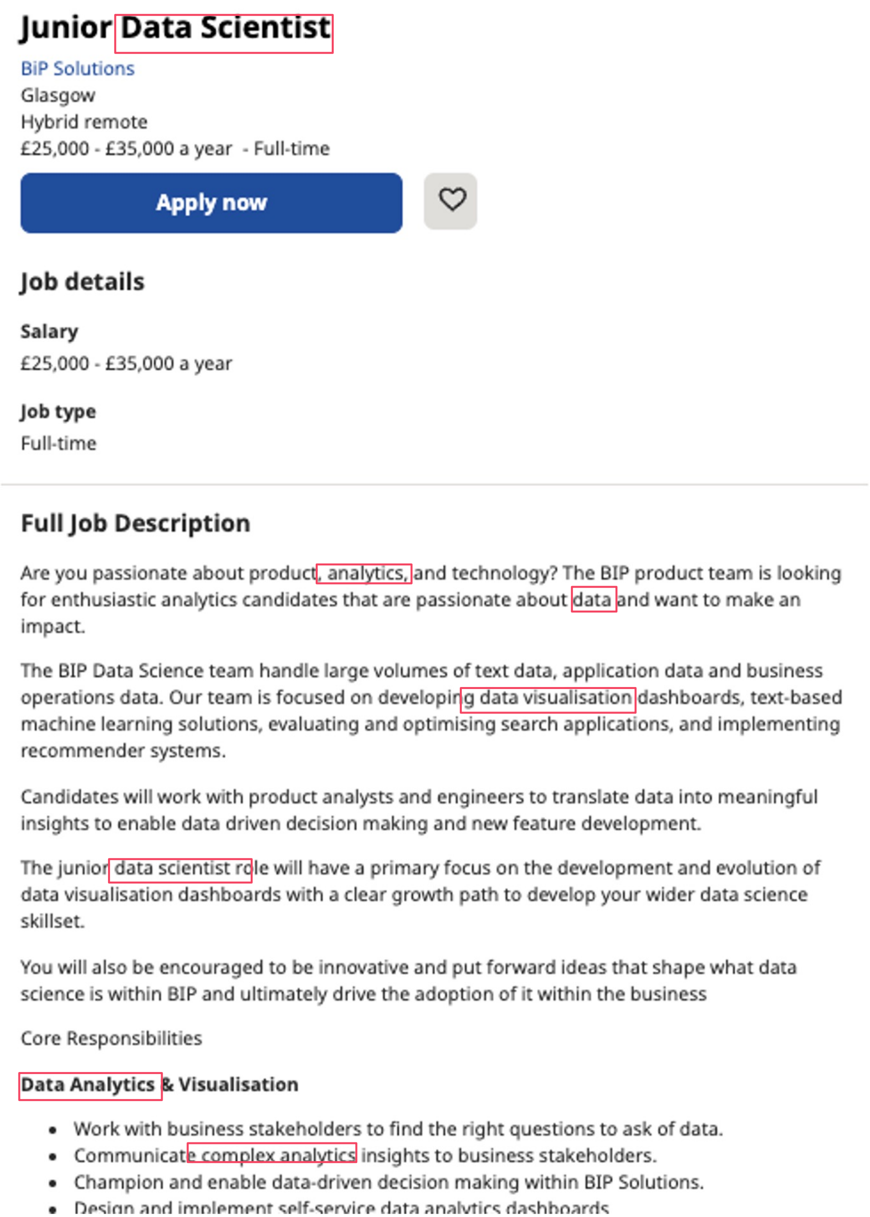
Core Responsibilities

Data Analytics & Visualisation

- Work with business stakeholders to find the right questions to ask of data.
- Communicate complex analytics insights to business stakeholders.
- Champion and enable data-driven decision making within BIP Solutions.
- Design and implement self-service data analytics dashboards



How do you (a human) classify a job posting in an occupation?



Junior Data Scientist

BiP Solutions
Glasgow
Hybrid remote
£25,000 - £35,000 a year - Full-time

Apply now

Job details

Salary
£25,000 - £35,000 a year

Job type
Full-time

Full Job Description

Are you passionate about product, analytics, and technology? The BIP product team is looking for enthusiastic analytics candidates that are passionate about data and want to make an impact.

The BIP Data Science team handle large volumes of text data, application data and business operations data. Our team is focused on developing data visualisation dashboards, text-based machine learning solutions, evaluating and optimising search applications, and implementing recommender systems.

Candidates will work with product analysts and engineers to translate data into meaningful insights to enable data driven decision making and new feature development.

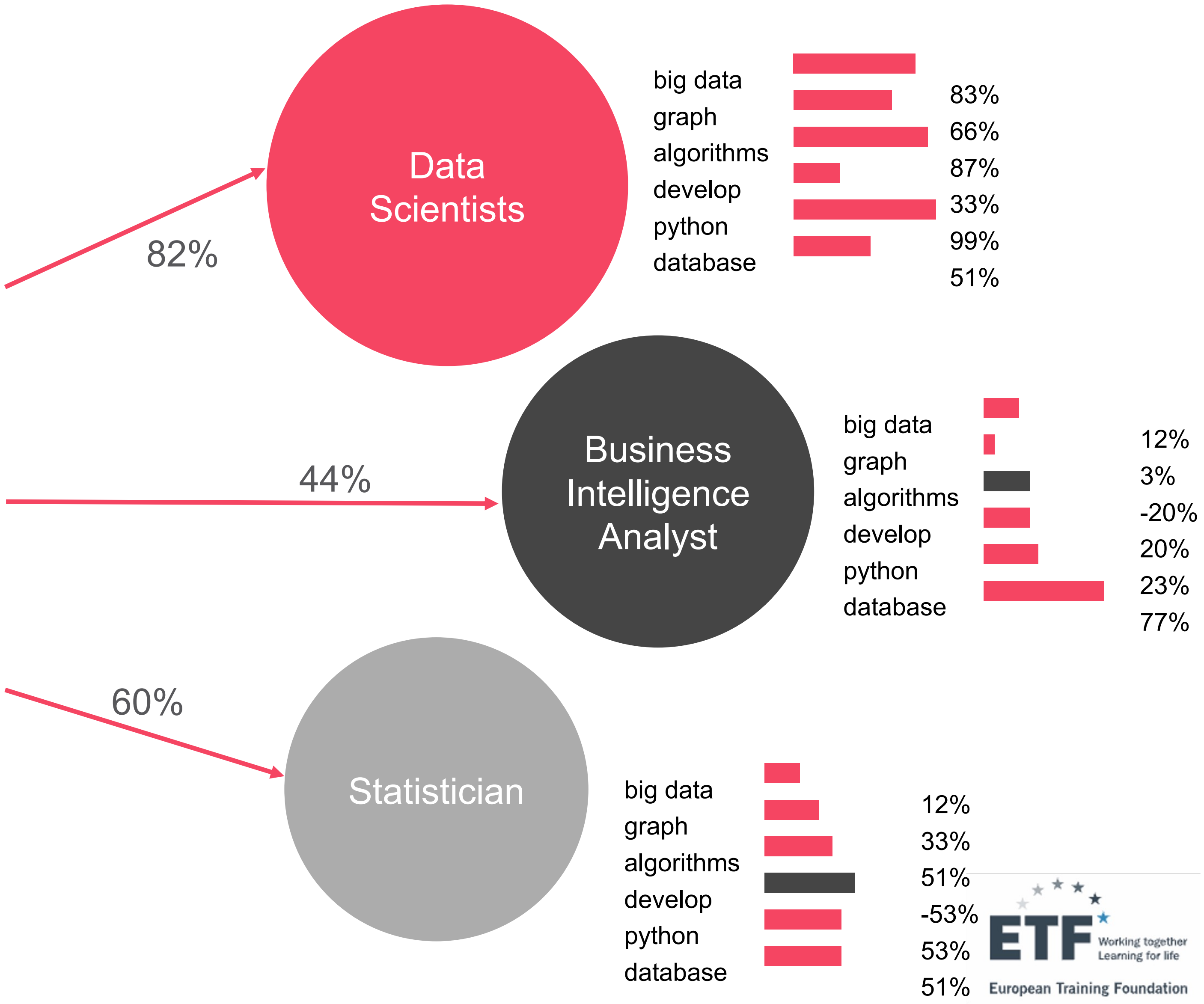
The junior data scientist role will have a primary focus on the development and evolution of data visualisation dashboards with a clear growth path to develop your wider data science skillset.

You will also be encouraged to be innovative and put forward ideas that shape what data science is within BIP and ultimately drive the adoption of it within the business

Core Responsibilities

Data Analytics & Visualisation

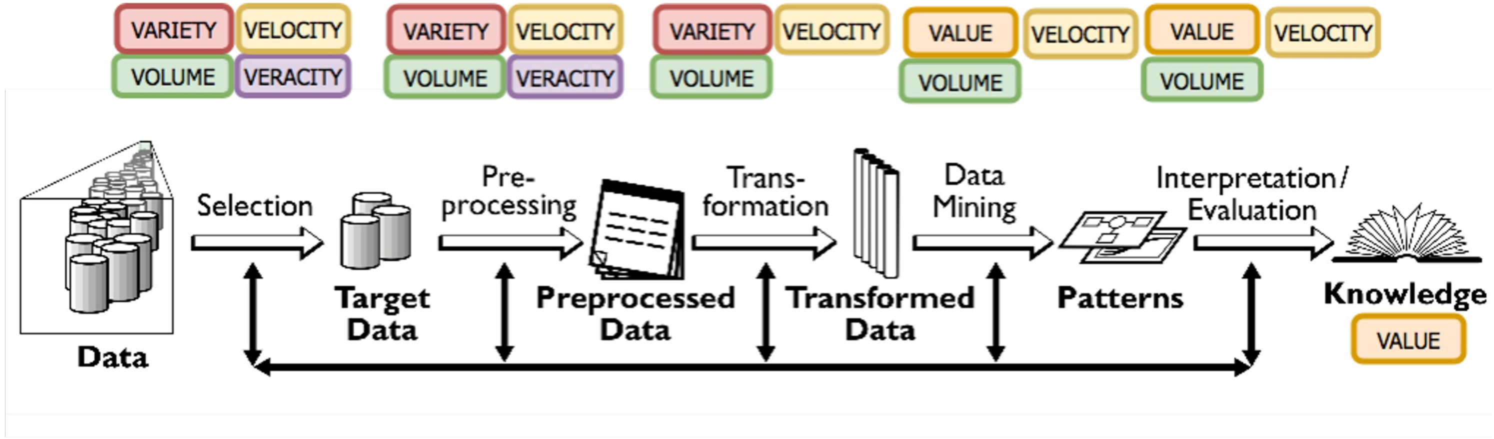
- Work with business stakeholders to find the right questions to ask of data.
- Communicate complex analytics insights to business stakeholders.
- Champion and enable data-driven decision making within BIP Solutions.
- Design and implement self-service data analytics dashboards



ETF Working together Learning for life
European Training Foundation

Methodological background

KDD – Fayyad, 1997



VARIETY

Unstructured data
(plain text to be processed)

Real time data

VELOCITY

Huge amount of data
(Terabytes)

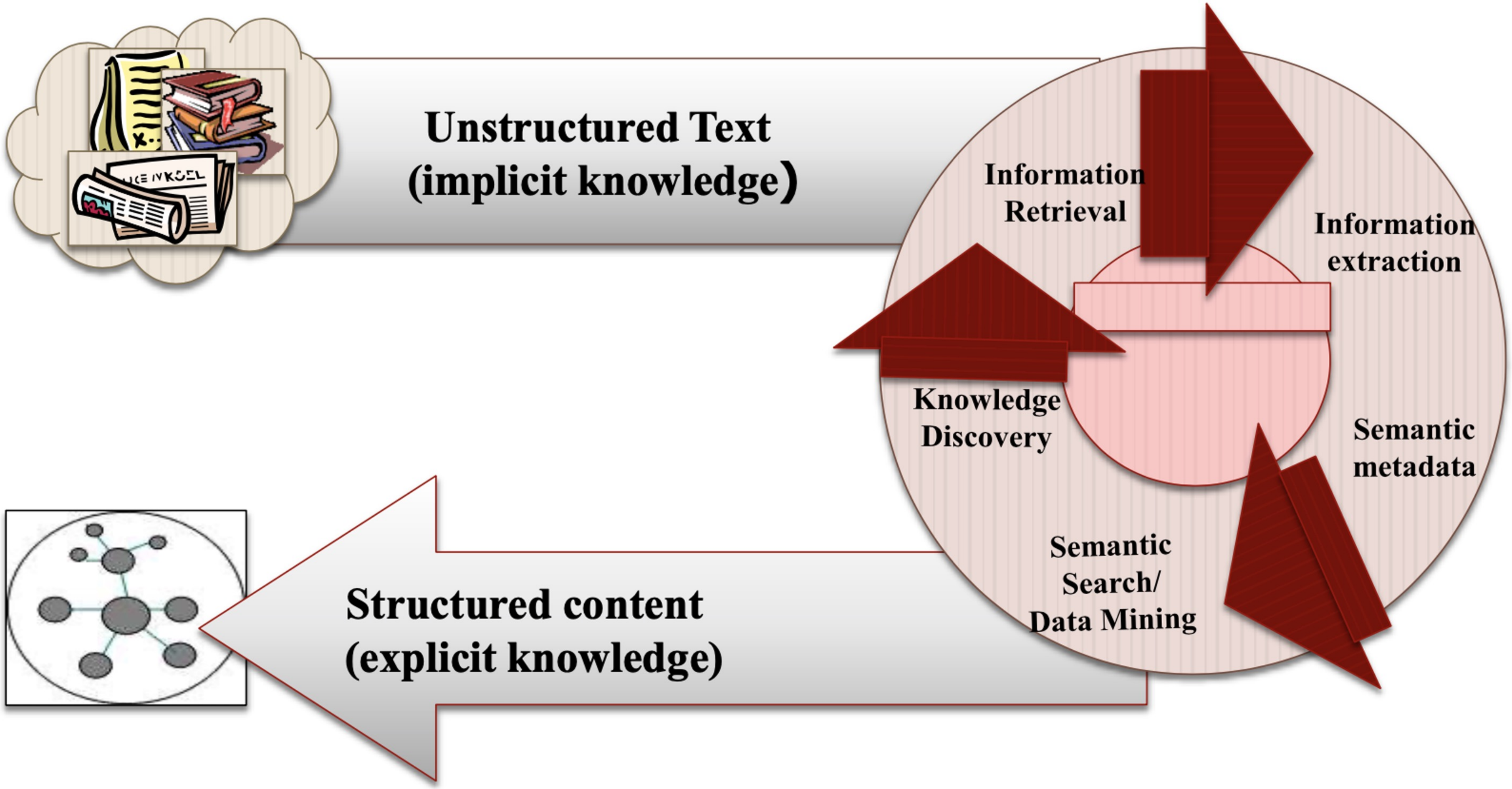
VOLUME

Data is noisy,
uncontrolled

VERACITY

Key components

- **Data ingestion**: **collect** raw data from OJV in both structured and unstructured (raw text) formats
- **Data processing**: **classify** data through **machine learning** techniques
- **Data analysis**: **extract information** from data and make it available through **visualization**



Unstructured Text
(implicit knowledge)

Information Retrieval

Information extraction

Semantic metadata

Semantic Search/ Data Mining

Knowledge Discovery

Structured content
(explicit knowledge)

Challenges

- Handle a **huge amount** of near real time data
- Data coming from web □ Need to detect and **reduce noise**
- **Multi language** environment
- Need to relate to **classification standards**
- Find a way to **summarize and present** a wide and complex scenario

What’s we need? The toolkit

Statistical Methods

Tools

User Experience Research

Time series analysis

Data mining

Missing data imputations

Multilevel modeling

Classification and clustering

Pattern recognition

AB testing

Principal component and factor analysis

Machine learning

Forecasting

Network analysis

Regression techniques

Big Data for LMI Summary

New sources

- Official statistics are representative and robust, but can lack detail and timeliness
- We need more ore frequently updated, fresh data
- We need more granular data to capture the real demand

Big Data For LMI

- Data derived from web job postings is the answer
- Up to date, detailed, adherent to reality
- Unstructured data, we can decode the DNA of the occupations by observing the skills required

System Overview and methodology

Lightcast

Topics

1. Stakeholders
2. The functional architecture
3. Data ingestion techniques
4. Data processing pipeline
5. Classification techniques

Stakeholders

Stakeholders

Project
Leader

Key
Users

Domain
Experts

End
Users

Project leader

- ETF
    - Lead the project with the steering committee
    - Define the scope of the project
    - Define key organizations
    - Maintain relations with EU stakeholders
    - Provide advice

Key Users

ETF, Lightcast, CRISP/University of Milan Bicocca

- Define requirements
- Monitor quality of the project
- Provide input to the development of the project
- Manage the landscaping
- Validate overall data flow and methodology

Domain Experts

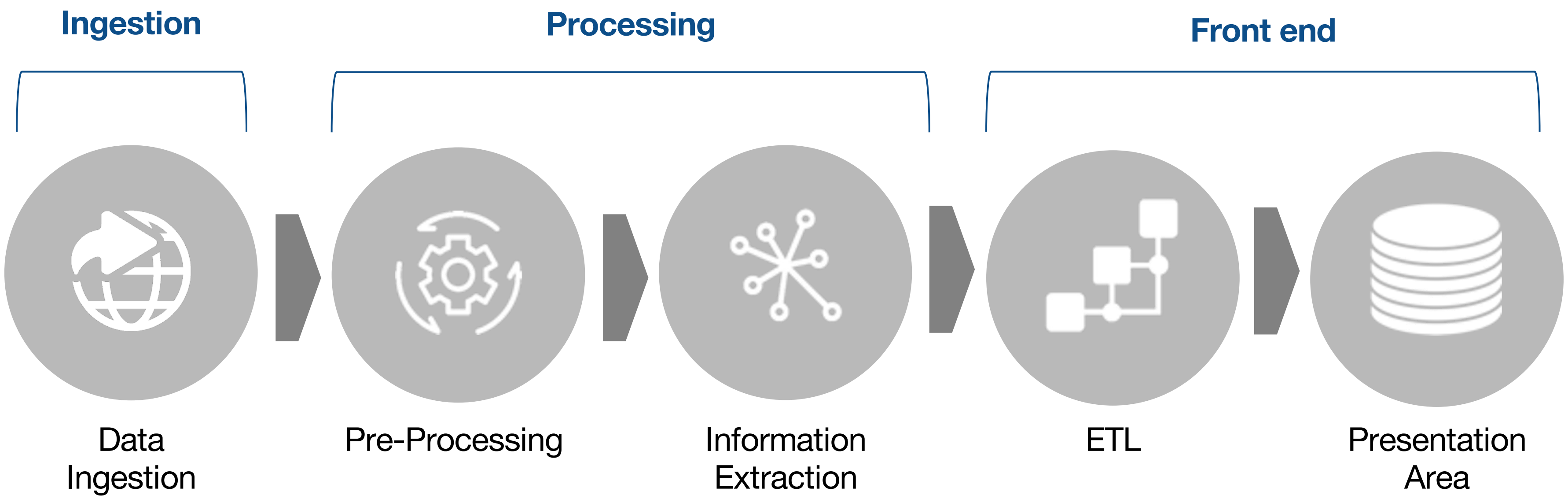
- International Country Experts
  - Provide the knowledge and expertise
  - Execute the landscaping
  - Understand the language/terms of their context
  - Evaluate the accuracy of the results
  - Test the product
  - Provide feedback

End Users

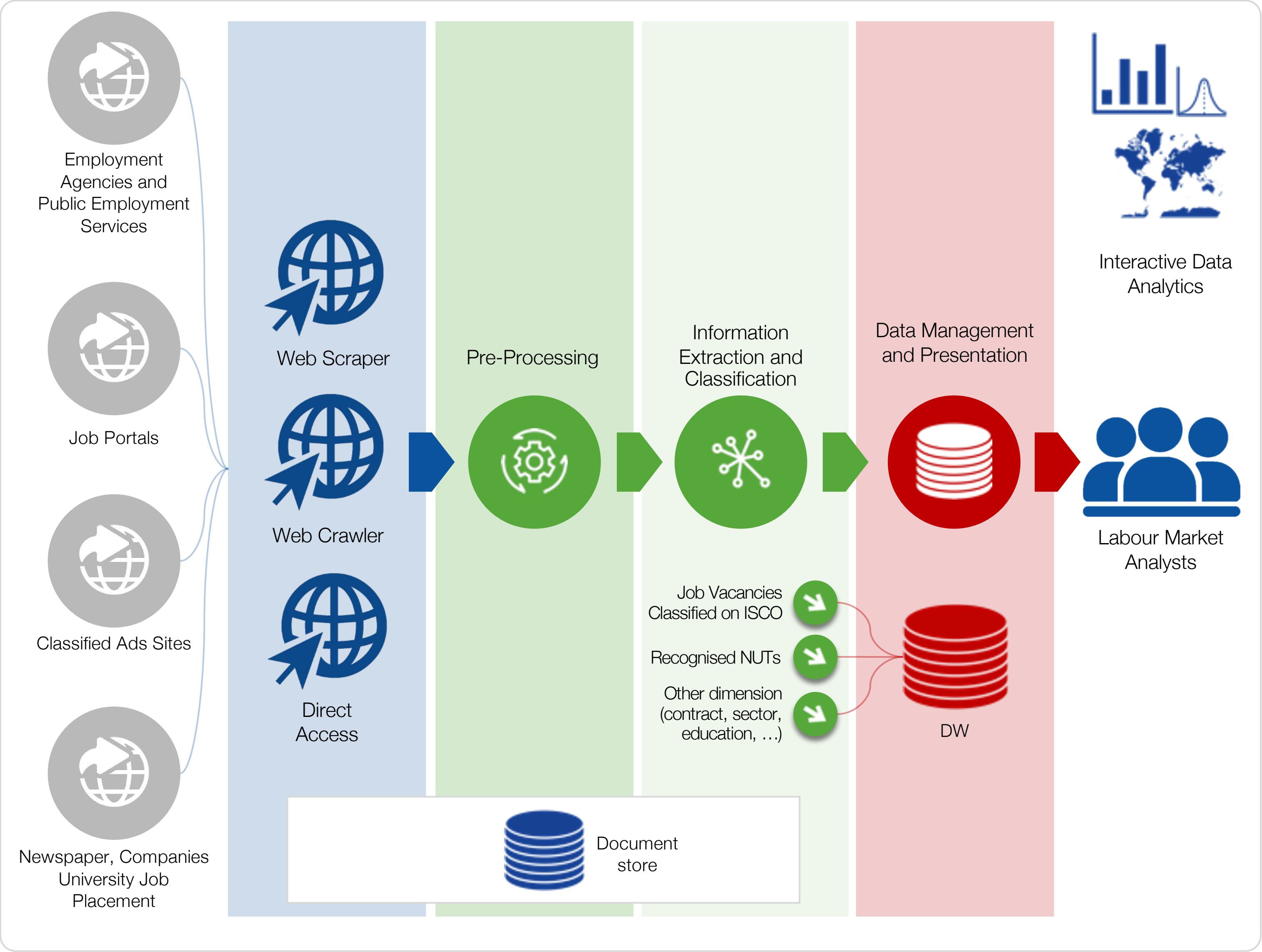
- Decision Makers and Business Users
    - (Visual) Explore dataset, analysis and aggregate data
    - Define new analysis processes
    - Produce Data storytelling
    - Make decisions by exploring data
- Data Scientists
    - Apply new machine learning models and AI techniques
    - Extract new insights from the data
    - Apply advanced data modelling to the dataset
- Data Analysts
    - Interprets data and turns it into information
    - Identifying patterns and trends
    - Extract and analyze aggregate data
    - Publish and share their analysis

The functional architecture

Overall Data Flow



Logical view


Employment Agencies and Public Employment Services
Job Portals
Classified Ads Sites
Newspaper, Companies University Job Placement
Web Scraper
Web Crawler
Direct Access
Pre-Processing
Information Extraction and Classification
Data Management and Presentation
Interactive Data Analytics
Labour Market Analysts
Job Vacancies Classified on ISCO
Recognised NUTs
Other dimension (contract, sector, education, …)
DW
Document store

Infrastructure Challenges

- Manage multiple parallel ingestion activities
- Availability of high performance computational infrastructure at a glance
- High memory requirements
- High storage volumes to store source and staging data
- Big data environment
- Scalable architecture

Data ingestion techniques

Landscaping

A **Landscaping activity** is performed to produce a list of **sources** (web portals) that are relevant for the Web Labour Market in a given country.

A Country Expert **validates** this list, that will become the initial step of the LMI System

Source selection strategy

4 Processing Steps

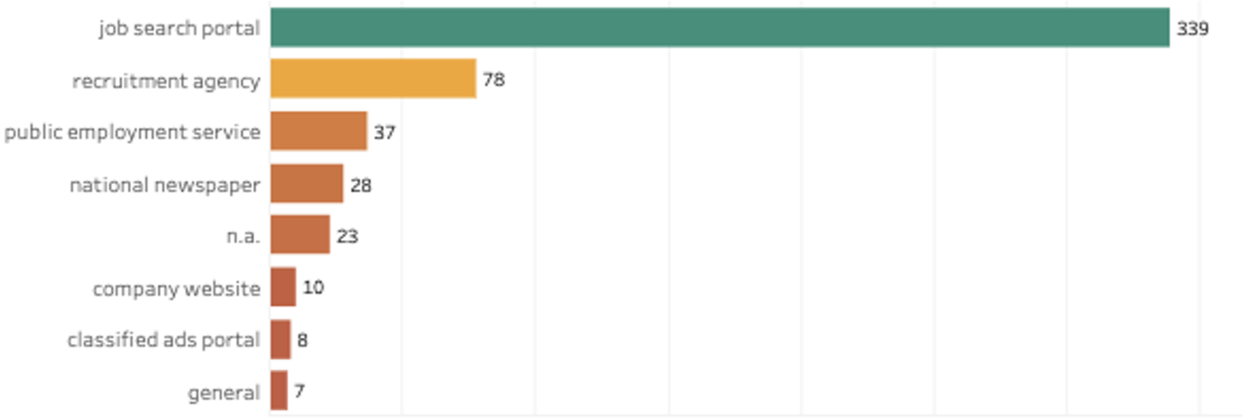
Source selection in landscaping

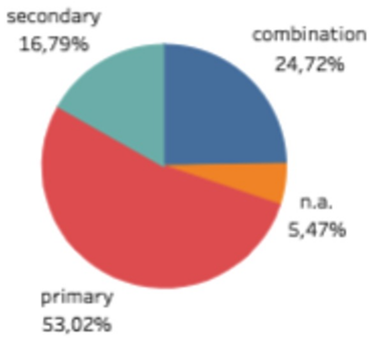
Augmentation

Agreements

Coverage

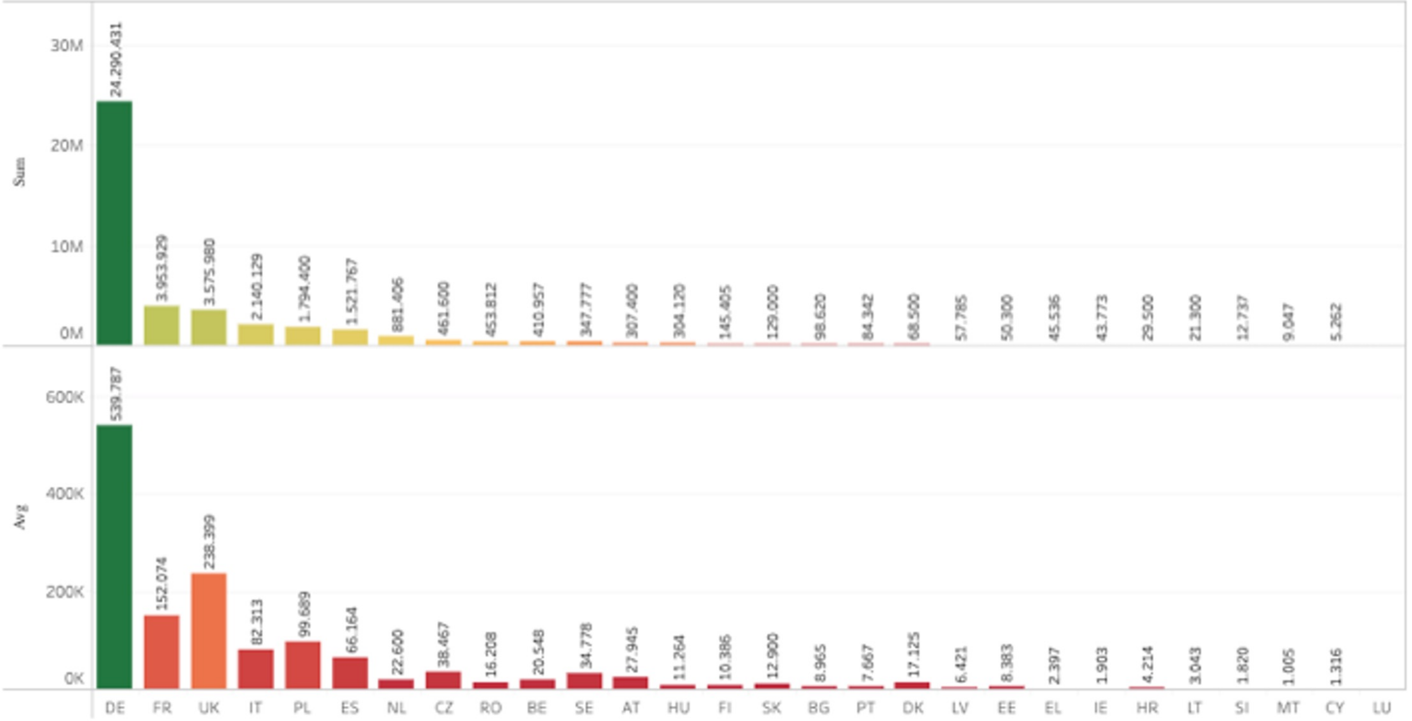
Sites by type of operator

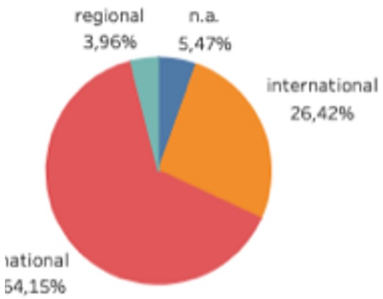


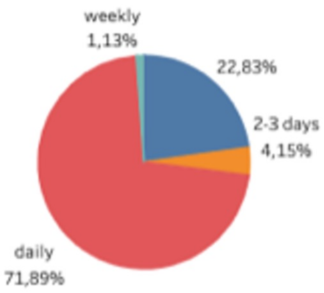


Vacancy volume by country

(estimated by ICE)







ETF Working together Learning for life

European Training Foundation

Relevance and ranking of sources

Volume

Type of
web portal

Data
Update

Structured
Data

Data Ingestion phase

The process of obtaining and importing data from web portals and storing them in a Database

Focus on volumes

Coverage augmentation & maximization

Direct agreements with the most relevant sources

Ingestion Challenges

Robustness of the process

Quality of data collected

Scalability and Governance

Ingestion Challenges

1. Robustness

Issue: potential technical problems when gathering data from a source (unavailability, block, changes in data structure)

Risk: loss of data

Solution: redundancy

- Have the most important sites (by volume and/or coverage) ingested from two or more sources
- Avoid loss of data in case of troubles with a source
- Collect data from both primary and secondary sources

Ingestion Challenges

2. Quality

Issue: need to obtain data as clean as possible, detecting structured data when available

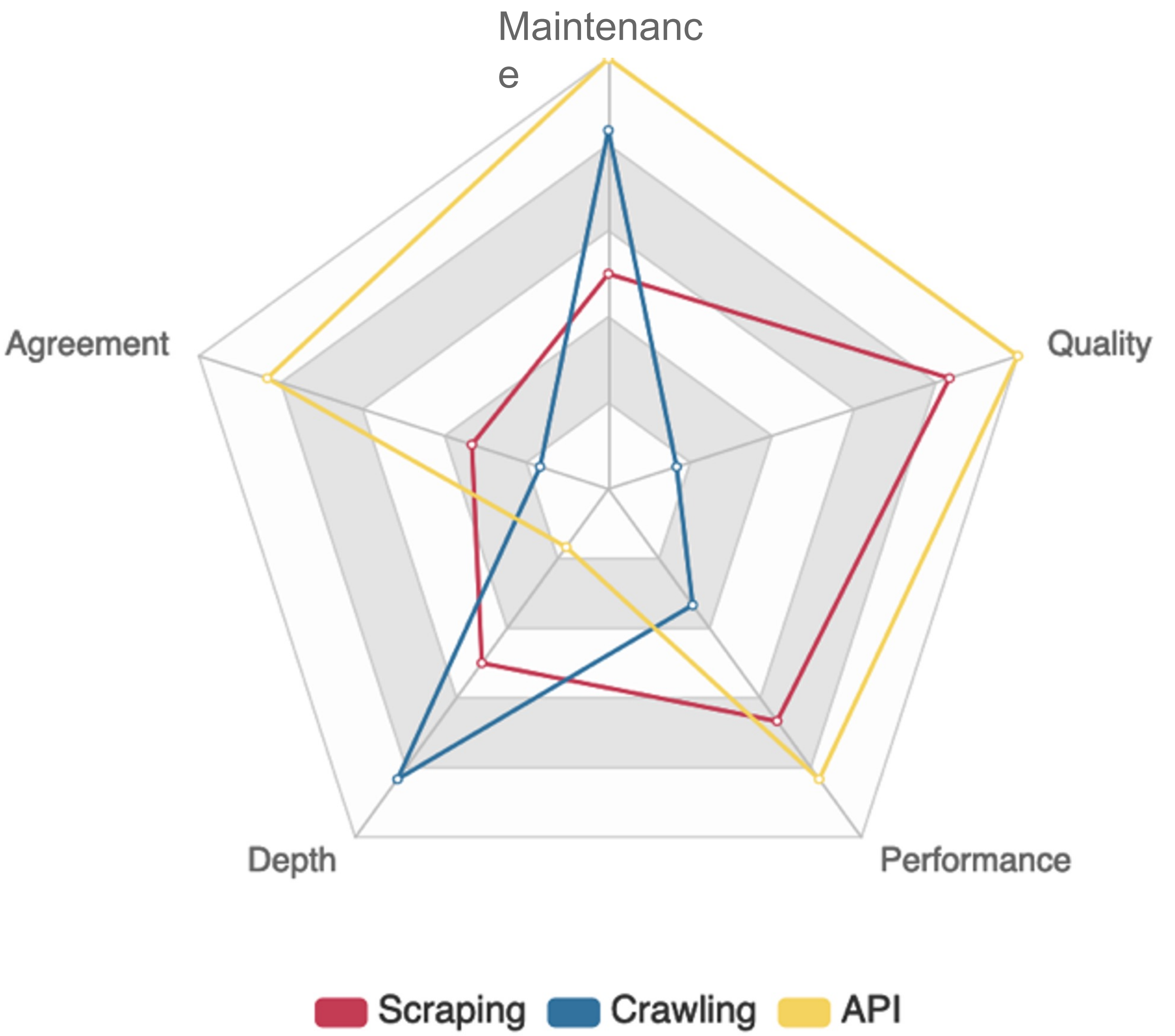
Risk: loss of quality

Solution: tailored ingestion. We collect data using a specific approach based on the single source:

- API
- Scraping
- Crawling

Ingestion Challenges - Quality

- API: when available (agreements), we collect mostly structured data from Web Portals.
  - Pros: Very high quality (most of fields structured)
  - Cons: Need agreement, not always available
- Scraping: if API is not feasible and the structure of the web poral is consistent, we develop a custom scraper that extract structured/unstructured data from pages
  - Pros: High Quality (many structured fields)
  - Cons: Web portal specific development
- Crawling: if web portal page structure is not consistent, we ingest data using a multi-purpose crawling approach
  - Pros: Lower quality (no structured fields)
  - Cons: Fast and Versatile approach


Maintenanc
e
Quality
Performance
Depth
Agreement
🟥 Scraping
🟦 Crawling
🟨 API

Ingestion Challenges

3. Scalability and Governance

Issue: need to handle a real and complex Big Data environment, simultaneously connecting to thousands of websites

Risk: Loss of Process control and loss of OJVs due to slowness of the process

Solution:

- ○ A scalable infrastructure
- ○ A monitoring and governance custom tool

Data processing pipeline

Data Pre-Processing – Challenges & Definitions

- 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 Pre-Processing – Challenges & Definitions

The process of cleaning ingested data and dedupicating OJVs, to guarantee that analytical phase'll work on data at the highest quality possible

Language detection

Noise reduction

OJVs Deduplication

Pre-Processing steps

Merging

Cleaning

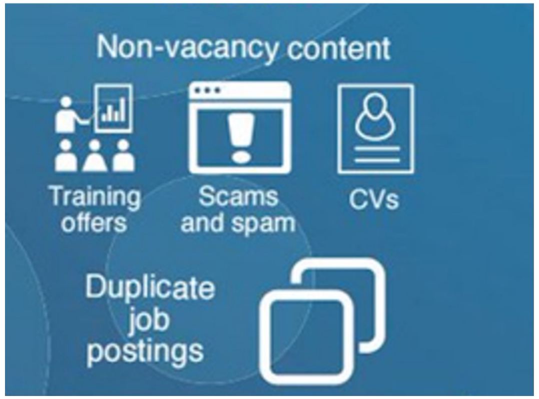
Text processing and summarizing

Data Pre-Processing
The language detection

- Why:
  - Each language has different keywords, stopwords,…
  - It can reflect different cultures and Labour Market scenarios…
  - … So it’s fundamental to classify the language of the OJV, so use the most proper classification pipeline
- How:
  - We trained for each language (60+) a specific classifier based on Wikipedia corpus
  - Obtained models are very accurate (~99% of precision) and fast to adopt in the pipeline
- What we obtain:
  - A fast and strong classification of the language used in each OJV
  - A way to archive OJVs for which we don’t have a classification pipeline

Data Pre-Processing
How to deal with noise?

- In a Big Data environment, we must deal with noise
  - Why? Because information in gathered from the web, one of the most noisy place ever known
- First of all, we've to master which type of noise we have to face with…:
  - Web pages explicitly not related to OJVs:
    - Social network pages
    - News pages
    - Privacy policy pages
    - …
  - Web pages disguised as OJVs:
    - Training courses
    - CVs
    - Consulting services
    - …



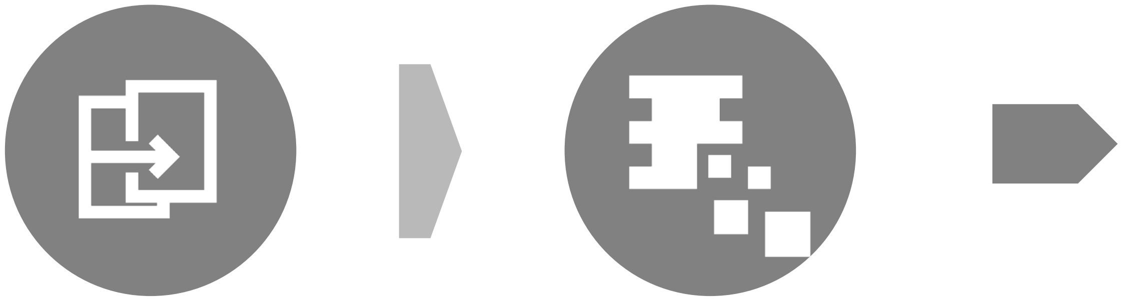
- …Then, we have to detect and handle duplicated OJVs:
  - Generally, a vacancy is posted on multiple portals
  - If we deal with them as distinct, we would overestimate Labour Demand
  - So, we've to detect duplicated OJVs and merge information coming from them in a single one

Data Pre-Processing
Noise Detection – How?

- 2 Steps approach:
  - Machine Learning approach
    - For each language, we trained a Naïve Bayes classifier with more than 20k web pages:
      - 10k of real OJVs related pages
      - 10k of web pages not related to OJVs
    - Accuracy of ~99%
    - Fast to train and use
    - An approach similar to a “Email Spam Detection” system
  - Fuzzy matching approach
    - Used to detect “OVJs like” webpages, but related to training offers, consulting services,….
    - It works looking ad page header and body to detect keywords (language dependent) that can help us label it like a “not-related to OJVs” page

But, before starting OJVs deduplication phase, we need to clean text to simplify and consolidate it…

Data Pre-Processing
Deduplication phase



Job ads

Physical deduplication or fuzzy matching

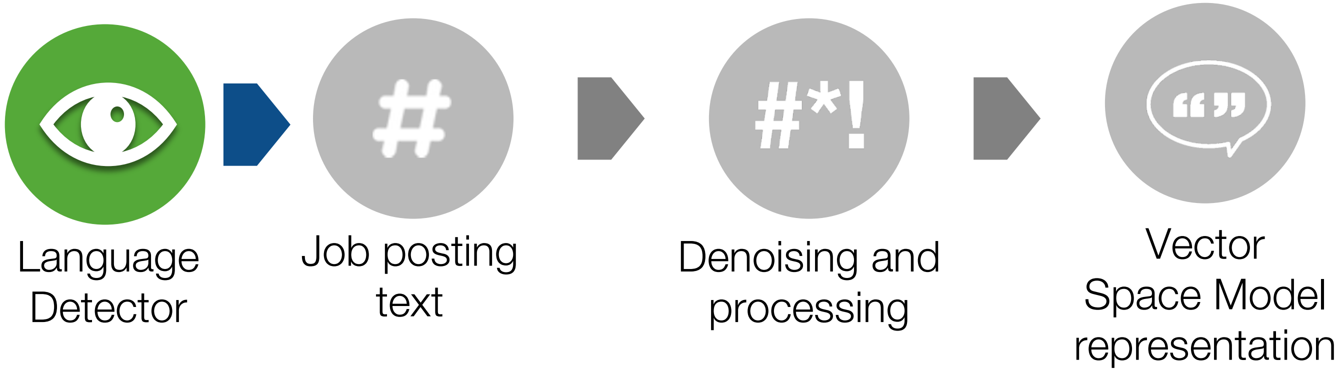
Made on the description (or content) part of the job vacancy.

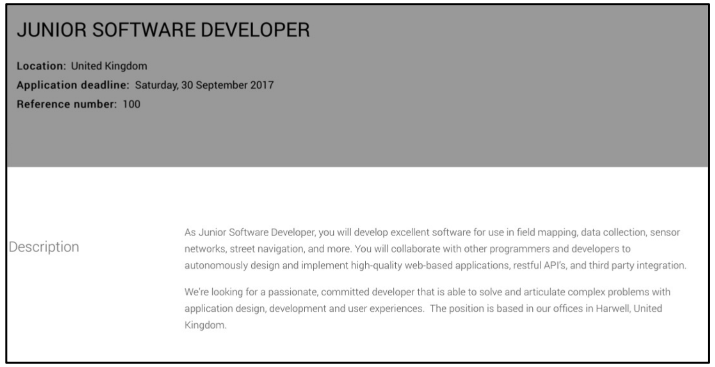
Metadata matching

Using metadata coming from job portals to remove job vacancies duplicates on the aggregators websites (e.g. reference id, page url)

Text processing and summarizing

The text processing and summarizing phase aims at reducing the text to improve the process of classifications of job vacancies according to the European standards.





JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom

Application deadline: Saturday, 30 September 2017

Reference number: 100

Description

As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful API's, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

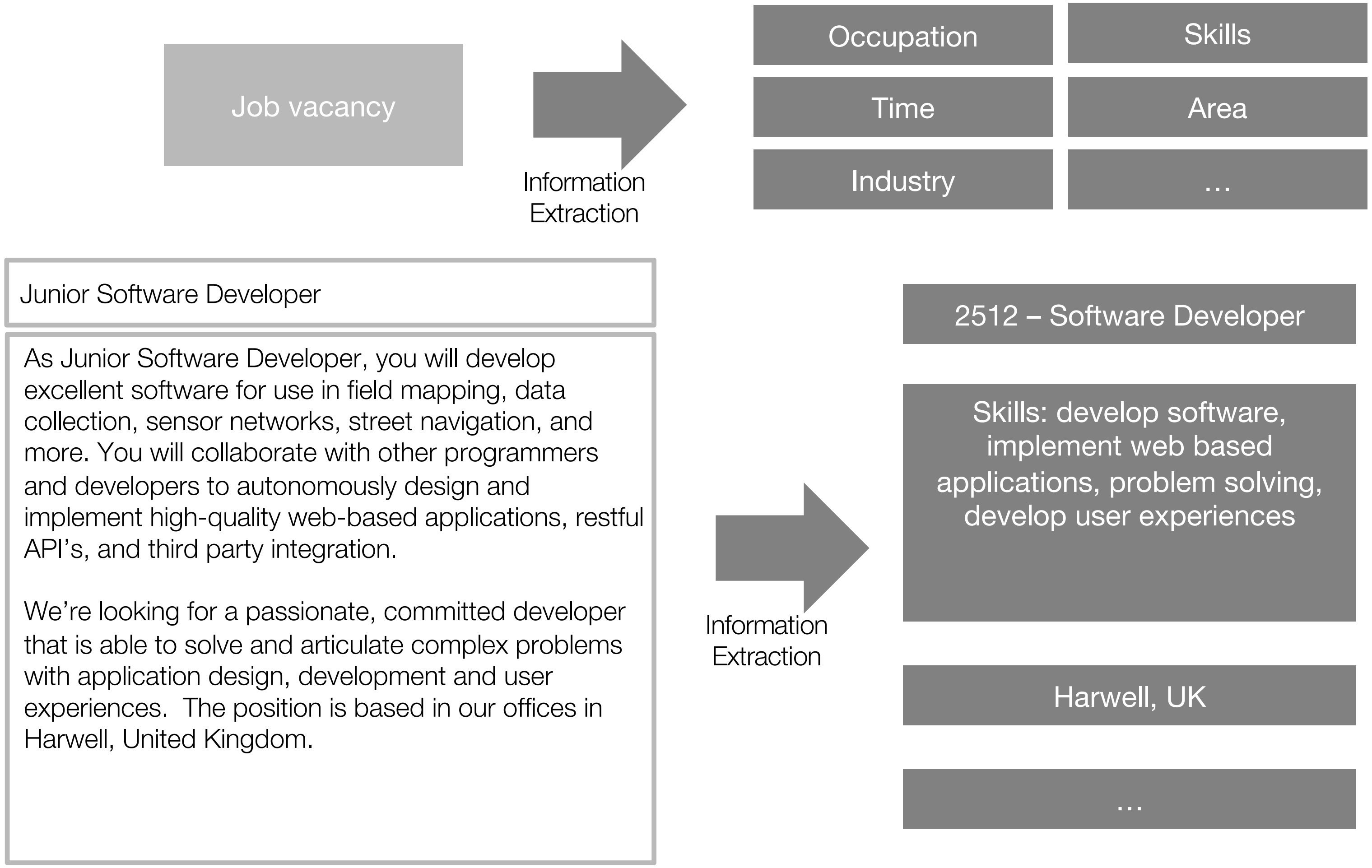
As Junior ⟨Software Developer⟩, you will develop excellent ⟨software⟩ for use in ⟨field mapping⟩, ⟨data collection⟩, ⟨sensor networks⟩, ⟨street navigation⟩, and more. You will ⟨collaborate⟩ with other ⟨programmers⟩ and ⟨developers⟩ to ⟨autonomously⟩ design and implement high-quality ⟨web-based applications⟩, restful ⟨API⟩'s, and third party ⟨integration⟩.

We're looking for a passionate, committed ⟨developer⟩ that is able to ⟨solve⟩ and articulate ⟨complex problems⟩ with ⟨application design⟩, ⟨development⟩ and ⟨user experiences⟩.
The position is based in our offices in ⟨Harwell⟩, ⟨United Kingdom⟩.

Classification techniques

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.



Job vacancy

Information Extraction

Occupation

Skills

Time

Area

Industry

…

Junior Software Developer

As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful API's, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

Information Extraction

2512 – Software Developer

Skills: develop software, implement web based applications, problem solving, develop user experiences

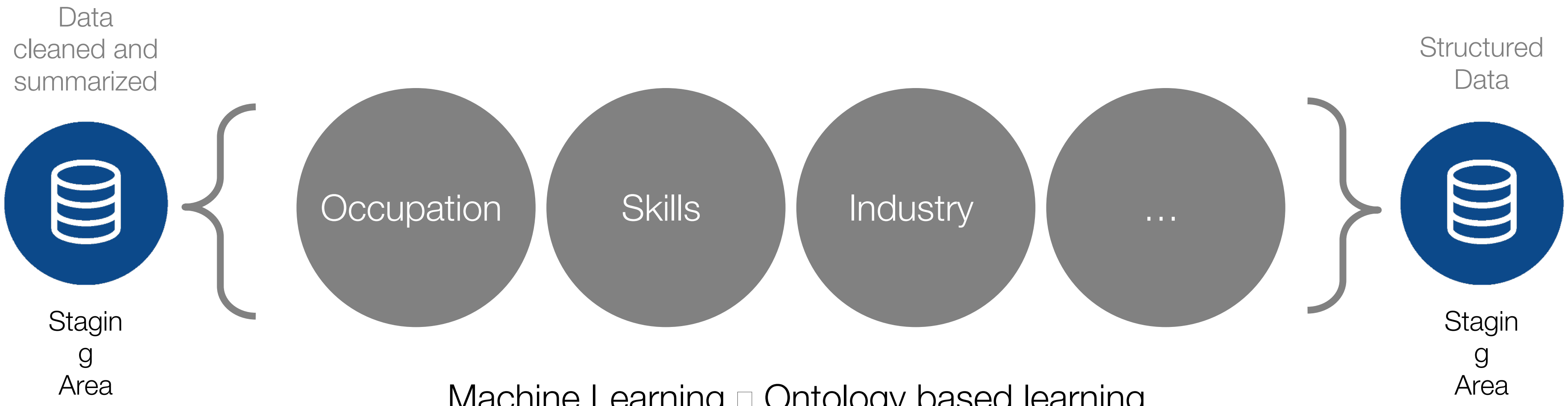
Harwell, UK

…

Information Extraction and Classification
Real Time Labour Market Intelligence

Information Extraction is an area of natural language processing that deals with finding factual information in free text.

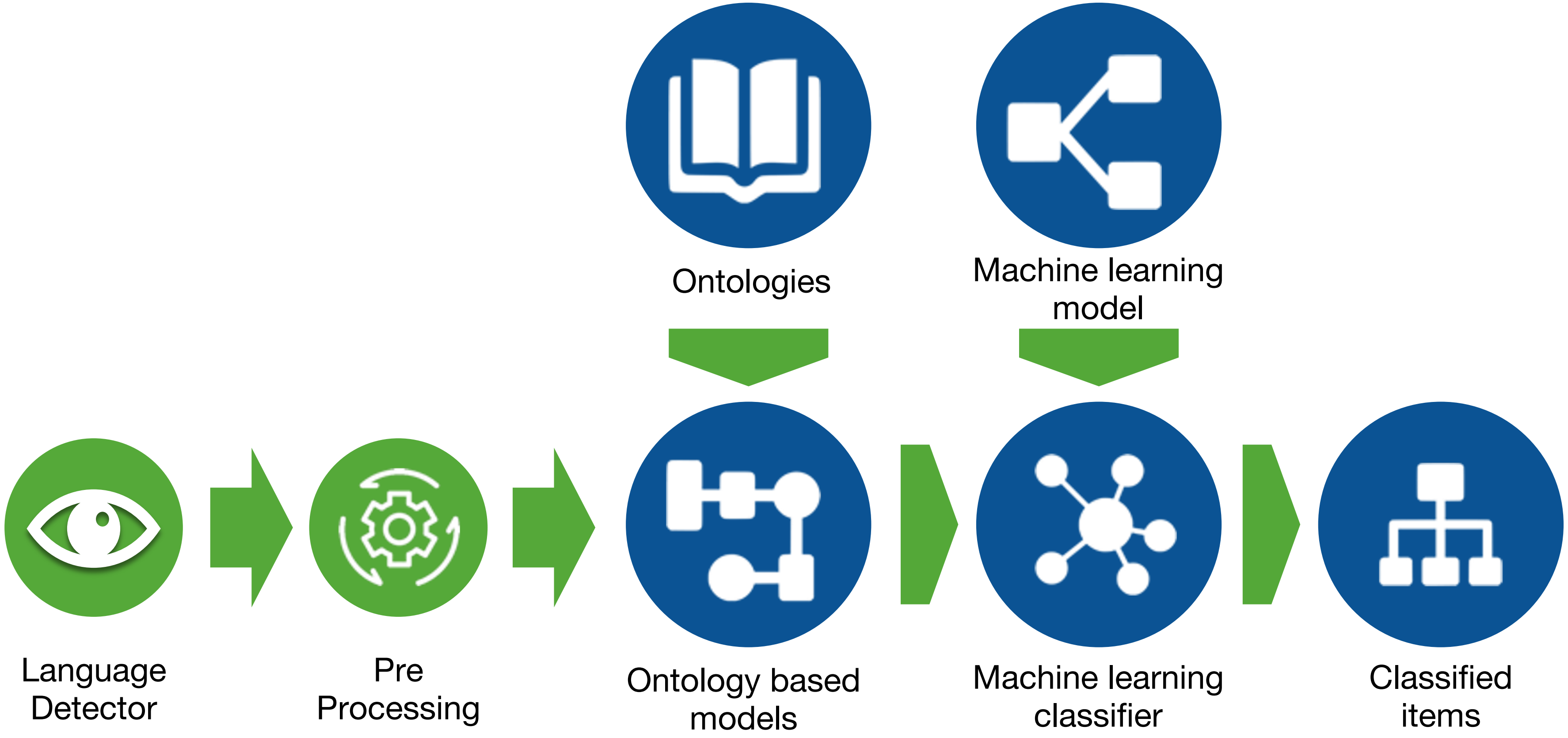
This task uses machine learning techniques (ontology based learning, supervised learning and unsupervised learning) to match job ads with standard classifications.



Machine Learning □ Ontology based learning, supervised learning and unsupervised learning, etc.

ETF Working together Learning for life
European Training Foundation

Occupations pipeline



Ontologies

Machine learning model

Language Detector

Pre Processing

Ontology based models

Machine learning classifier

Classified items

ETF Working together Learning for life

European Training Foundation

Text Similarity Approaches

String based

String similarity measures operate on string sequences and character composition.

Jaro-Winkler, Jaccard, Cosine similarity

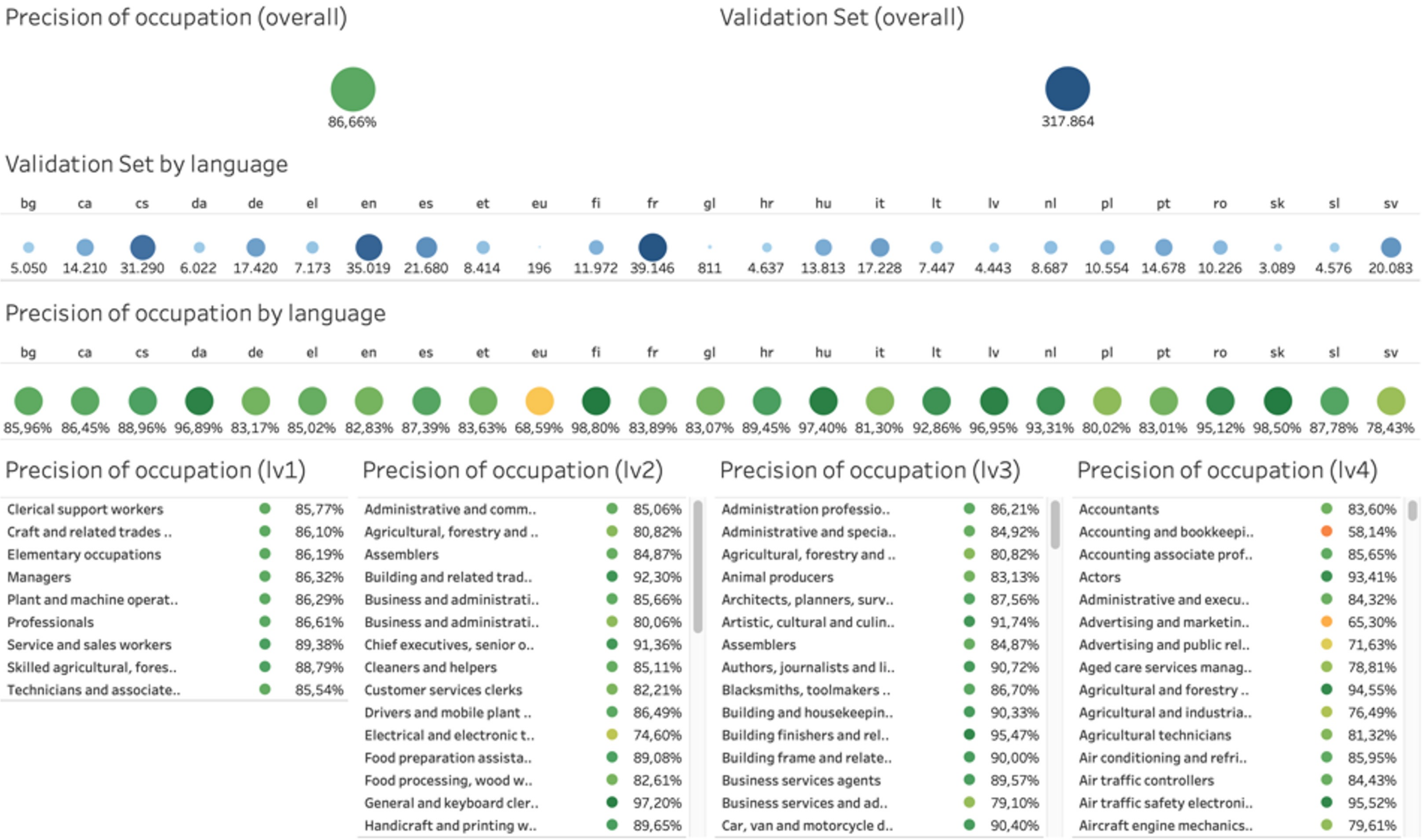
Corpus based

Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora.

Latent Semantic Analysis, Explicit Semantic Analysis, DIStributionally similar words using CO-occurrences

Knowledge based

Knowledge-Based Similarity is based on identifying the degree of similarity between words using information derived from semantic networks



Precision of occupation (overall)

86,66%

Validation Set (overall)

317.864

Validation Set by language

bg	ca	cs	da	de	el	en	es	et	eu	fi	fr	gl	hr	hu	it	lt	lv	nl	pl	pt	ro	sk	sl	sv
5.050	14.210	31.290	6.022	17.420	7.173	35.019	21.680	8.414	196	11.972	39.146	811	4.637	13.813	17.228	7.447	4.443	8.687	10.554	14.678	10.226	3.089	4.576	20.083

Precision of occupation by language

bg	ca	cs	da	de	el	en	es	et	eu	fi	fr	gl	hr	hu	it	lt	lv	nl	pl	pt	ro	sk	sl	sv
85,96%	86,45%	88,96%	96,89%	83,17%	85,02%	82,83%	87,39%	83,63%	68,59%	98,80%	83,89%	83,07%	89,45%	97,40%	81,30%	92,86%	96,95%	93,31%	80,02%	83,01%	95,12%	98,50%	87,78%	78,43%

Precision of occupation (lv1)

Clerical support workers	●	85,77%
Craft and related trades ..	●	86,10%
Elementary occupations	●	86,19%
Managers	●	86,32%
Plant and machine operat..	●	86,29%
Professionals	●	86,61%
Service and sales workers	●	89,38%
Skilled agricultural, fores..	●	88,79%
Technicians and associate..	●	85,54%

Precision of occupation (lv2)

Administrative and comm..	●	85,06%
Agricultural, forestry and ..	●	80,82%
Assemblers	●	84,87%
Building and related trad..	●	92,30%
Business and administrati..	●	85,66%
Business and administrati..	●	80,06%
Chief executives, senior o..	●	91,36%
Cleaners and helpers	●	85,11%
Customer services clerks	●	82,21%
Drivers and mobile plant ..	●	86,49%
Electrical and electronic t..	●	74,60%
Food preparation assista..	●	89,08%
Food processing, wood w..	●	82,61%
General and keyboard cler..	●	97,20%
Handicraft and printing w..	●	89,65%

Precision of occupation (lv3)

Administration professio..	●	86,21%
Administrative and specia..	●	84,92%
Agricultural, forestry and ..	●	80,82%
Animal producers	●	83,13%
Architects, planners, surv..	●	87,56%
Artistic, cultural and culin..	●	91,74%
Assemblers	●	84,87%
Authors, journalists and li..	●	90,72%
Blacksmiths, toolmakers ..	●	86,70%
Building and housekeepin..	●	90,33%
Building finishers and rel..	●	95,47%
Building frame and relate..	●	90,00%
Business services agents	●	89,57%
Business services and ad..	●	79,10%
Car, van and motorcycle d..	●	90,40%

Precision of occupation (lv4)

Accountants	●	83,60%
Accounting and bookkeepi..	●	58,14%
Accounting associate prof..	●	85,65%
Actors	●	93,41%
Administrative and execu..	●	84,32%
Advertising and marketin..	●	65,30%
Advertising and public rel..	●	71,63%
Aged care services manag..	●	78,81%
Agricultural and forestry ..	●	94,55%
Agricultural and industria..	●	76,49%
Agricultural technicians	●	81,32%
Air conditioning and refri..	●	85,95%
Air traffic controllers	●	84,43%
Air traffic safety electroni..	●	95,52%
Aircraft engine mechanics..	●	79,61%