Big Data for Labour Market Information

Session 3

Knowledge Discovery in Databases (KDD) for LMI

Alessandro Vaccarino – Fabio Mercorio

Big Data for Labour Market Information – focus on data from online job vacancies – training workshop

Milan, 21-22 November 2019

ETF Working together Learning for life

European Commission

EaP Eastern Partnership

Topics

1. KDD steps for game changing in LMI
2. What you can do with LMI processed? [examples]
3. Issues, limitations and challenges

Is Big Data a game changer in the field of labour market?

Three main Labour Market Sources can support LM Intelligence

(1) Statistical sources
(2) Administrative sources
(3) Web Sources (Big Data 4 LMI)

Quo vadis Labour Market?

LM CHALLENGING FACTORS

1. Skills Evolution
2. New Emerging Occupations
3. Job Automatisation/Replacement
4. Mobility

LM NEEDS

1. Updated information (near-real-time)
2. Data driven decisions (let data speak)
3. Prediction can be done to anticipate trends

Knowledge becomes crucial to support different LM actors and policy makers in understanding LM dynamics and trends

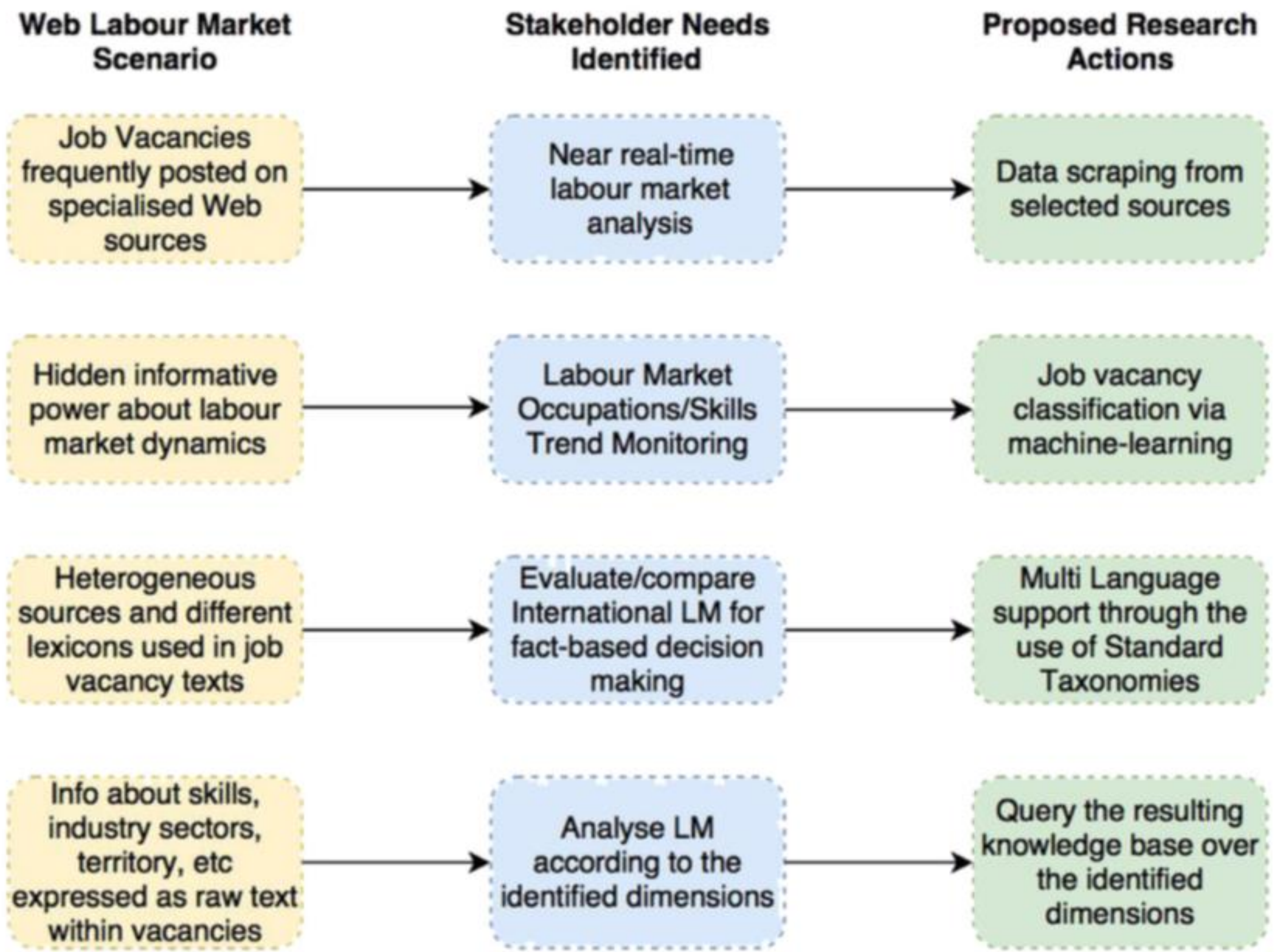

Web Labour Market Scenario
Stakeholder Needs Identified
Proposed Research Actions
Job Vacancies frequently posted on specialised Web sources
Near real-time labour market analysis
Data scraping from selected sources
Hidden informative power about labour market dynamics
Labour Market Occupations/Skills Trend Monitoring
Job vacancy classification via machine-learning
Heterogeneous sources and different lexicons used in job vacancy texts
Evaluate/compare International LM for fact-based decision making
Multi Language support through the use of Standard Taxonomies
Info about skills, industry sectors, territory, etc expressed as raw text within vacancies
Analyse LM according to the identified dimensions
Query the resulting knowledge base over the identified dimensions

Table 1 Main characteristics for LM Data Sources

LM Source Type	Data Type[2]	Generation Rate	Data Model Paradigm	Quality	Coverage	Analysis Paradigm	Believability	Value
Statistical	Structured	Periodically	Relational	Owner's responsibility	Owner's responsibility	Top Down & Model Based	Owner's responsibility	intrinsic
Administrative	Structured or Semi-structured	Periodically	Relational	Owner's responsibility	Owner's responsibility & User's responsibility	Top Down & Model Based	Owner's responsibility & User's responsibility	intrinsic
Web	Structured, Semi-structured or Unstructured	Near-real-time or real-time	Relational and Non Relational (NoSQL)	User's responsibility	User's responsibility	Bottom up & Data Driven	User's responsibility	extrinsic

Table 2 Most significant limitations of Big Data architectures

Issue (most significant)	Caused by	Conceptual Blocks of Big Data Architectures
Schema-free data are out: only structured data sources can be manipulated. Roughly, this means that only data that obey a rigid, well-defined data model can be handled, to the exclusion of all "unstructured" data, such as free text, comments and Web content in general.	Variety	Data ingestion; NoSQL models;
No adaptability to change: the addition of a new source requires the whole process to change, and this makes it difficult to scale the architecture over multiple (albeit structured) sources.	Variety, Velocity	Data lake
Rigid ETL: the procedures that transform content from source formats to target formats have to be precisely written to fit the desired data structure (e.g., data warehouse).	Variety	Schema free; data-driven approach (bottom-up rather than top-down)
Time consuming: the larger the volume of data to be processed, the greater the time needed to complete the process. ETL procedures are usually high time and memory consumers, as they need to "scan" all the data sources at any time to transform source data.	Volume, Variety, Velocity	Scale-out rather than scale-up

How to deal with OJVs at scale?

Web Job Vacancy example

Job Title: Data Scientist**.**

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

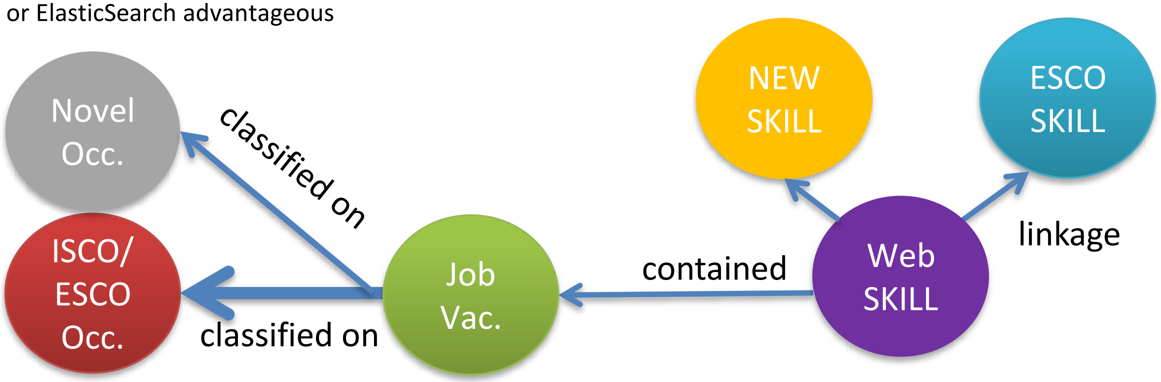
- SQL and relational databases;
- Data analysis with R (or Matlab);
- Processing large data sets with MapReduce and Hadoop);
- Real time analytics with Spark, Storm or similar;
- Machine Learning;
- Natural Language Processing (NLP) and text mining;
- Development in C++, Python, Perl;
- Experience with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

Web Job Vacancy example

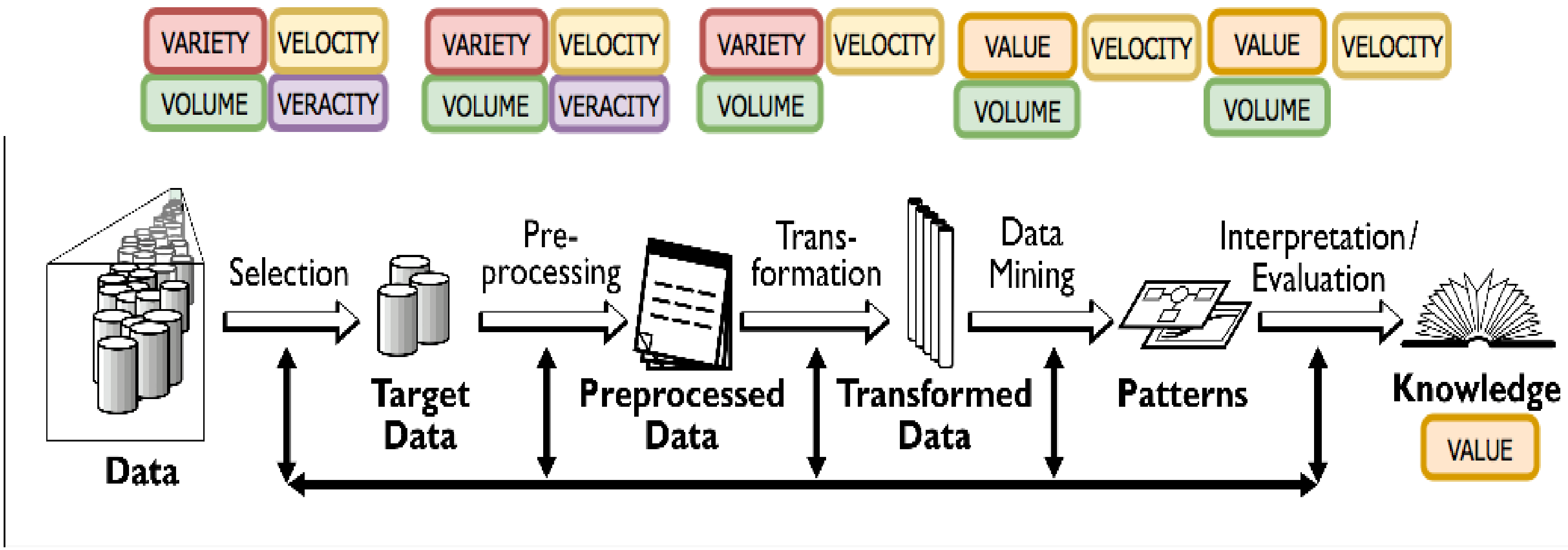
Job Title: Data Scientist**.**

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

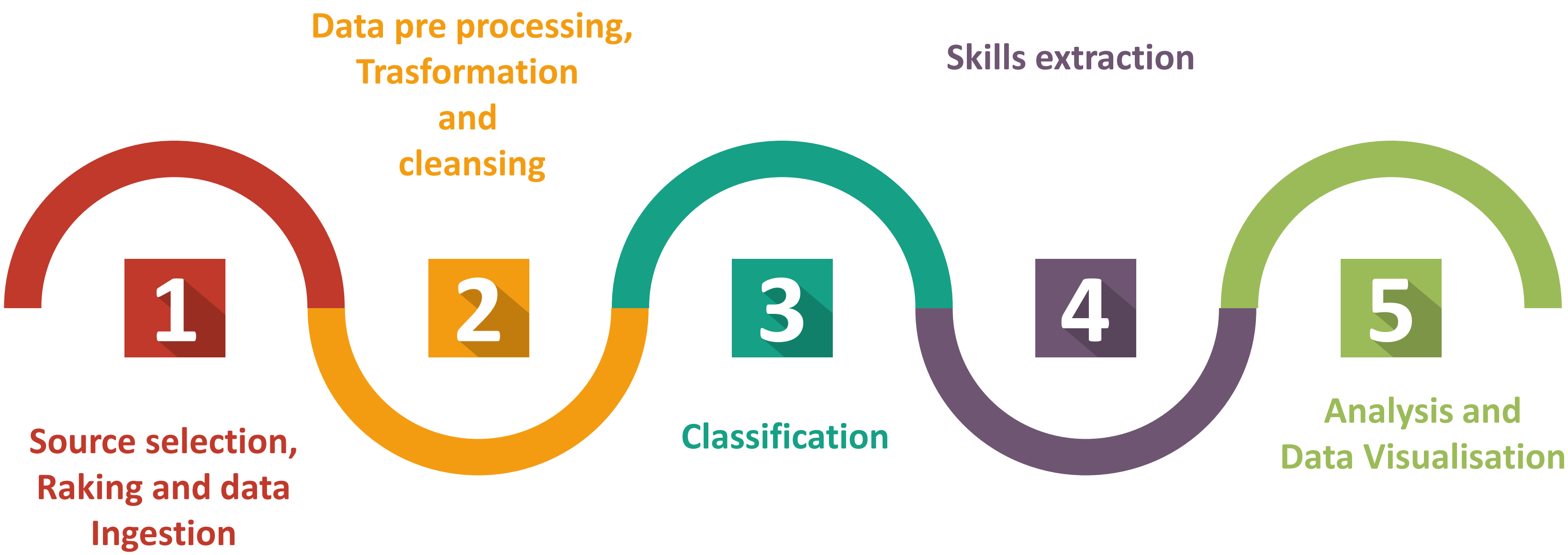
- SQL and relational databases;
- Data analysis with R (or Matlab);
- Processing large data sets with MapReduce and Hadoop);
- Real time analytics with Spark, Storm or similar;
- Machine Learning;
- Natural Language Processing (NLP) and text mining;
- Development in C++, Python, Perl;
- Experience with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



The Process (KDD, Fayyad 1997)



The Process (KDD 4 LMI)



Which expertises are needed to build a LMI system?

Five steps for turning Big Data into LMI

Professionals involved in the different steps of the LMI project:

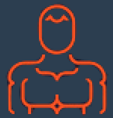
- Statisticians
- Data Engineer and Data scientists
- LM Domain experts – economists and sociologists

These expertise owned by such professionals show the necessity for a multidisciplinary approach in developing a LMI project

Five steps for turning Big Data into LMI

This step includes text processing, NLP, data cleaning, denoising, and deduplication

Statisticians : Identify measures of data quality, data distribution and significance



Computer scientist : Guarantee pre—processing to scale over million items

LM Domain experts : How do we identify LM domain synonyms that help in improving data accuracy? How do we identify criteria that characterise missing values and duplicates?

Five steps for turning Big Data into LMI

It includes data reduction and projection, which aim at identifying a unified model

Statisticians : Measure completeness of unified data model

Computer scientist : Guarantee transformation from raw data into target data at scale

LM Domain experts : How do we identify the destination data format and taxonomy ?

Five steps for turning Big Data into LMI

identify appropriate AI algorithms (e.g., classification, prediction, regression, clustering, information filtering), by searching for patterns of interest in a particular representational form, based on the purpose of the analysis

Statisticians & Computer scientists : Algorithm identification, parameters tuning, implementation

LM Domain experts : Which skills should be selected and which should be discarded?
Result evaluation

Five steps for turning Big Data into LMI

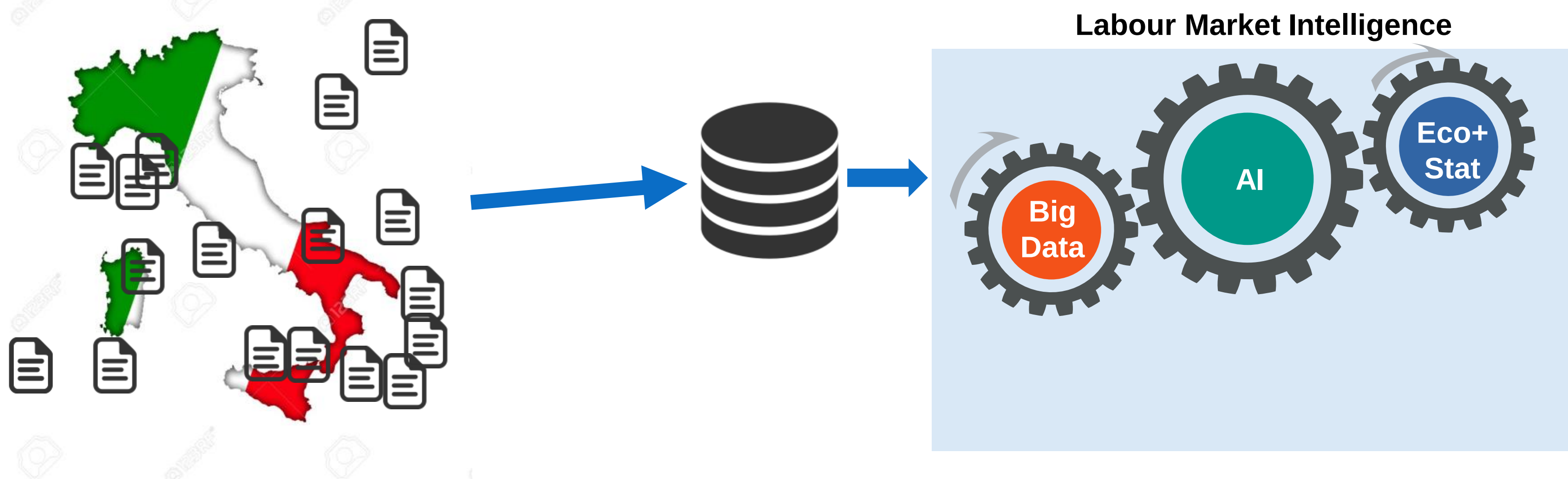
Employ visual paradigms to visually represent the knowledge obtained, depending on the user's objectives. In the LMI context, it means taking into account the user's ability to understand the data and their main goal

Statisticians & Computer scientists : Identify visualisation and narrative paradigm and implement it

LM Domain: How do we deliver appropriate knowledge according to stakeholder needs?
How do we identify visual navigation paths for each stakeholder?
How do we retrieve feedback (if any) from LM users?
How do we put LM knowledge into business?

ITALIAN Real-Time Labour Market Monitor

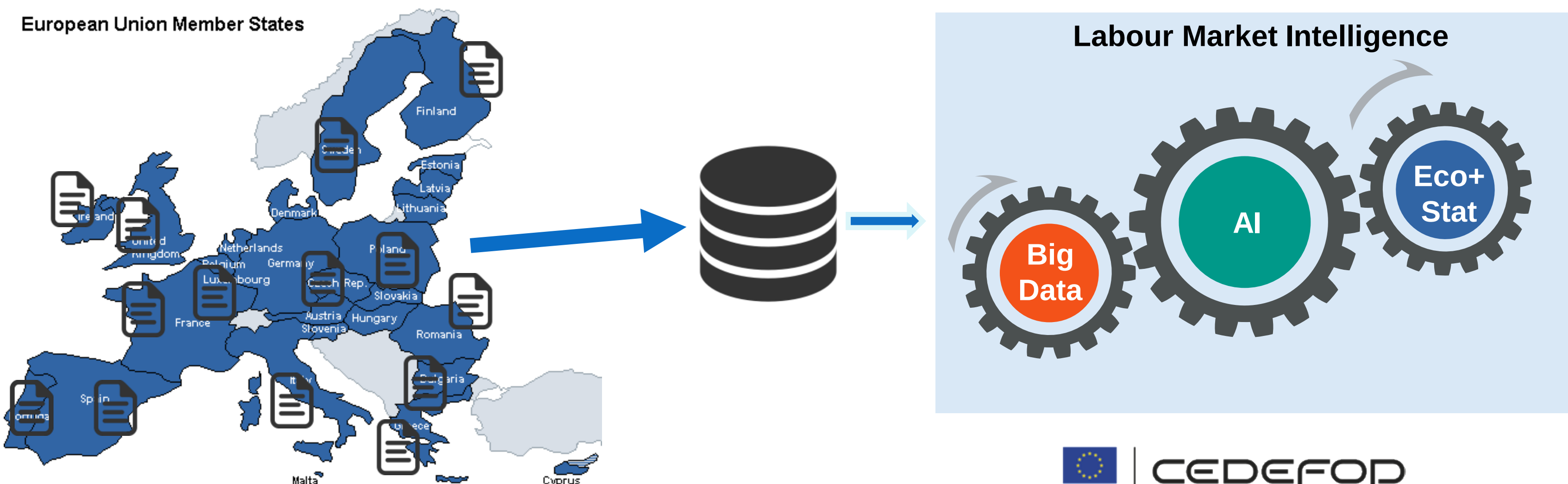
powered by Tabulaex



interuniversity research centre for public services

OJV since 2013 – 4M+ vacancies unique

EUROPEAN Real-Time Labour Market Monitor



CEDEFOP
European Centre
for the Development
of Vocational Training

28 EU Countries – 32 Languages – more than 6M unique vacancies per month

What you can do with LMI processed?

1. Occupation and Skill Discovery
2. Soft/Digital/Hard Skill Rates
3. New Emerging Occupations
4. Taxonomy Extension
5. Explain how ML works behind the scenes to humans

Occupation and Skill Discovery

Focus on occupations and skills requested by the online-LM

Live demo: Territorial dimension

VIDEOS

Soft/Digital/Hard Skill Rates

How to estimate the impact of digitalization within occupations?

Compute Skill Rates

Goal: Estimate the pervasiveness of ICT in both ICT and not ICT-related jobs

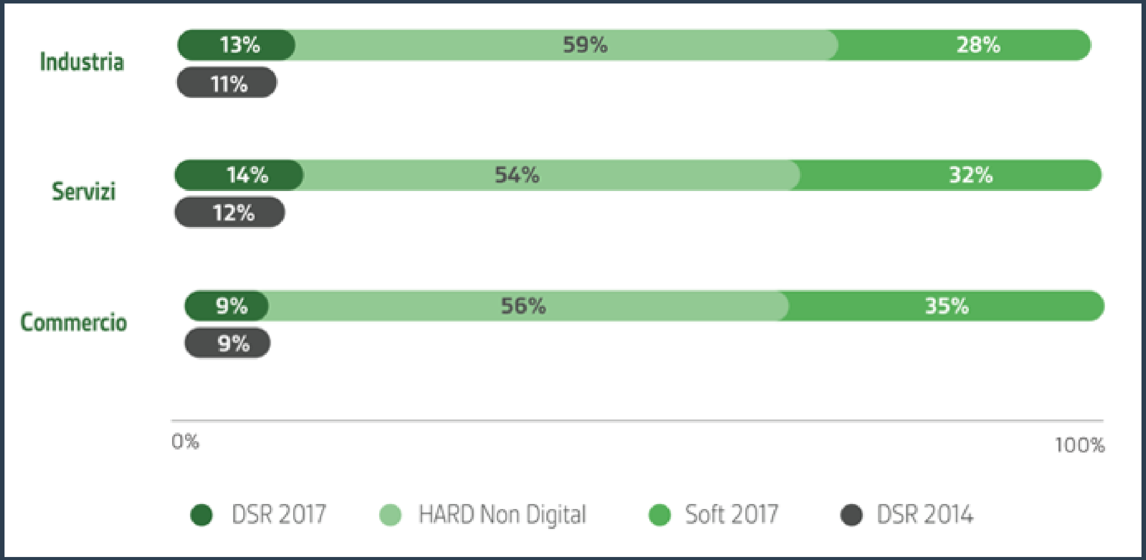
Idea: Exploit the informative power of Classified OJV for computing The Digita Skill Rate (DSR), Soft skill rate and Hard non digital Skill Rate

DSR estimates the incidence of digital skills in a single profession and comes from observing the pervasiveness of digital skills in all professions whether they are related to the ICT world or not.

Compute Skills Rates

Source WollyBI

Demand of digital, specialist and soft skills - by sector

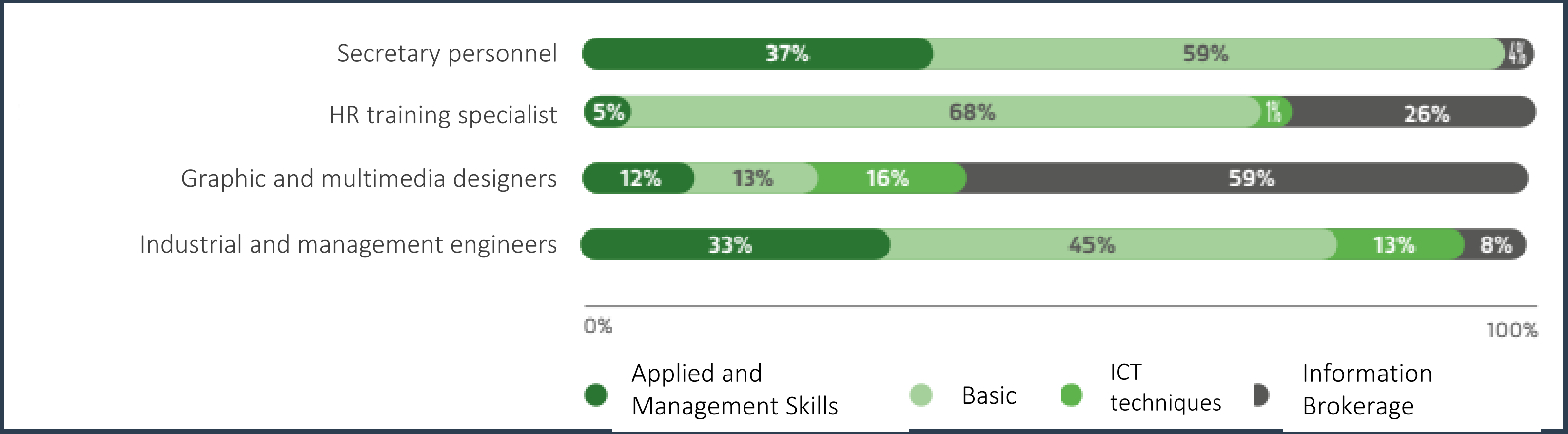


Ad hoc analyses at different level of granularity, here focusing on the «sector»….

Credits to WollyBI: a trademark of TabulaeX

Compute Skills Rates

Source WollyBI



- **Applied and Management Skills** = ability to use tools and software to manage both operational and decisional processes
- **ICT Techniques Skill**= very specialized on solutions, platforms and programming languages
- **Basic Skill** = for everyday use of basic IT tools
- **Information Brokerage Skill** = for the use of IT tools aimed at corporate communication

… and more, looking at each occupation…

Credits to WollyBI: a trademark of TabulaeX

Compute Skills Rates [ESCO skills + novel]

SOURCE WOLLYBI

	Applied and Management Skills			ICT techniques		Information Brokerage	
Occupation	Database usage	Programs for draughtsman	3D modelling	Front-end Website implementation	Web programming	Graphic Software Usage	SW markup usage
Graphic and multimedia designers	2.5	3.5	3	4.5	2	5	2

	Applied and Management Skills		Information Brokerage		
Occupation	Database usage	ERP	Digital data management	SEO Search Engine Optimiz.	Social Network Usage
HR training specialist	4.5	4	4.5	4	2.5

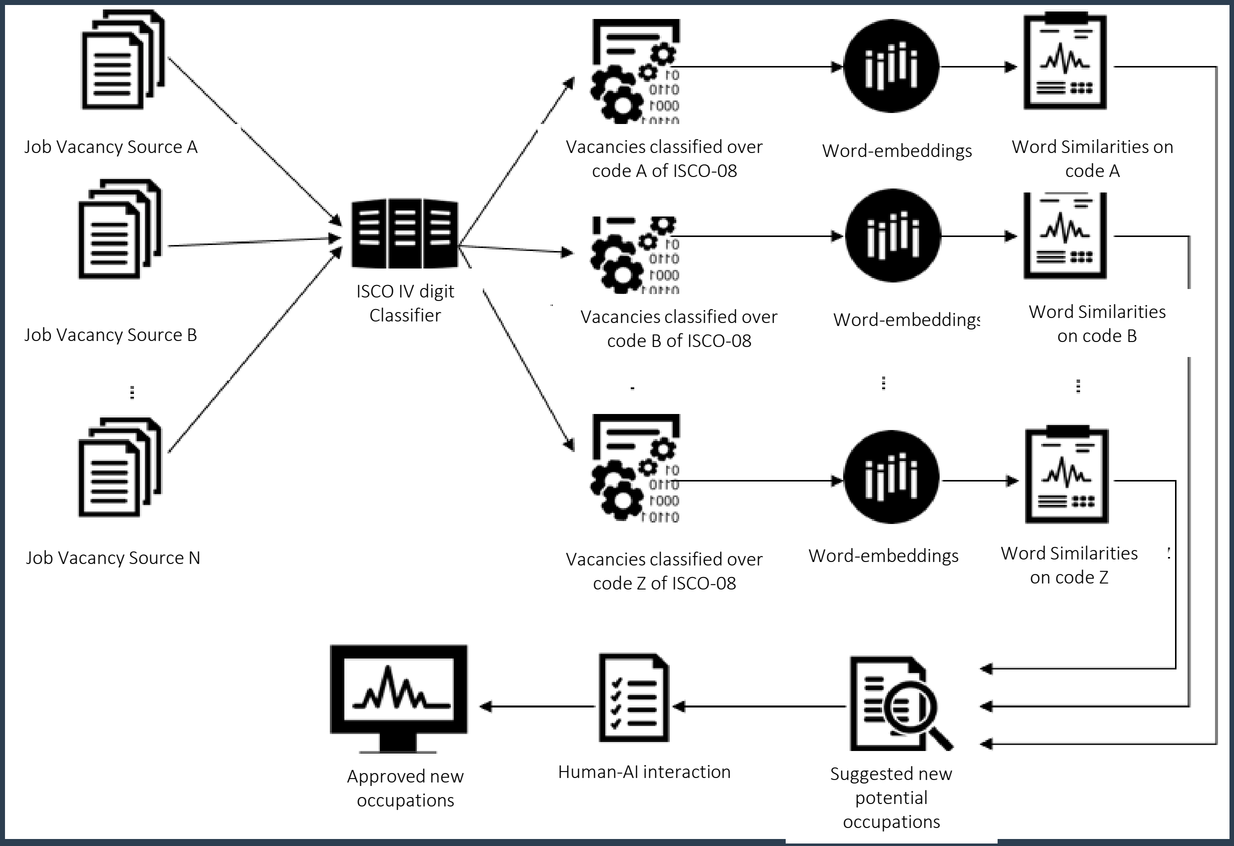
… and more, looking at elementary skills

Credits to WollyBI: a trademark of TabulaeX

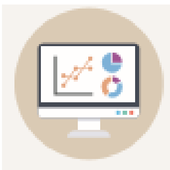
New Emerging Occupations on the basis of skill (dis)similarities

Detecting new emerging occupations through AI

1. Classify OJV over ISCO-iv digit
2. Build-up several vector-space representations of words (occupations and skills) to catch lexicon similarities between OJVs
3. Compute similarities between known terms (occupations and skills) and new ones
4. Suggest new potential occupations for Human-AI validation



(Some) New Emerging Occupations



Data Scientist

Cloud Computing

Cyber Security Expert

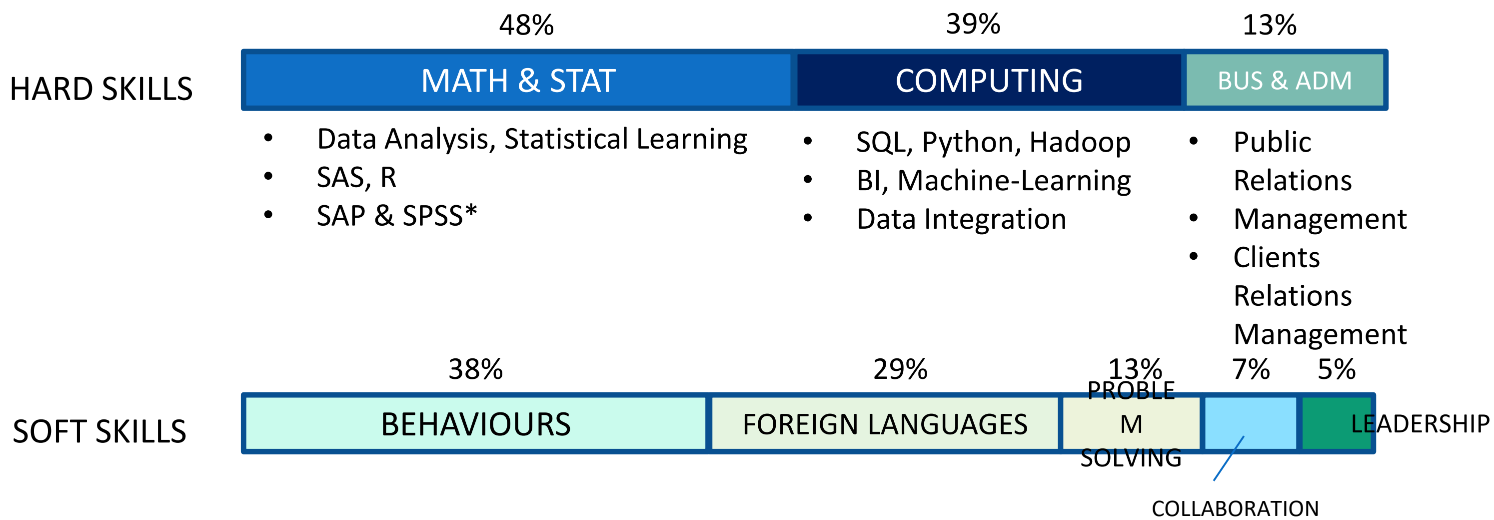
Business Intelligence Analyst

Big Data Analyst

Social Media Marketing

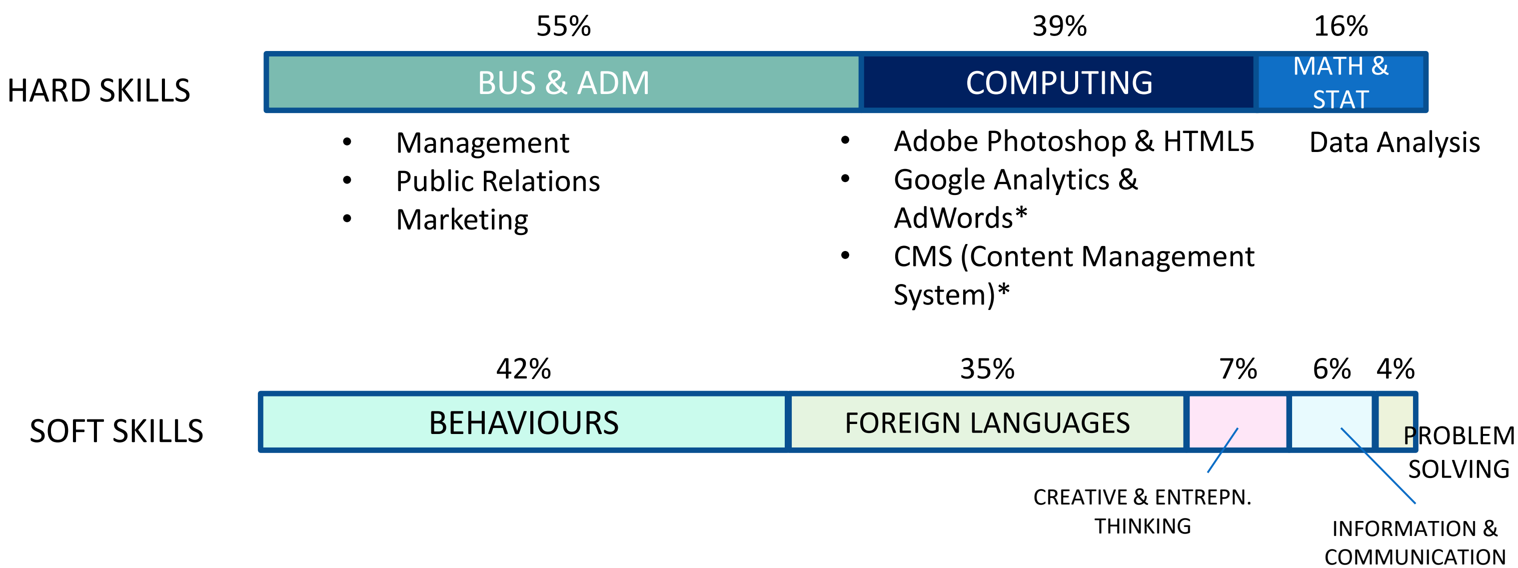
9,000 Web Job Vacancies collected related to (some) new emerging occupations (above) between Jan-2014 and August-2019

DATA SCIENTIST – 1,7k vacancies 2014-2019



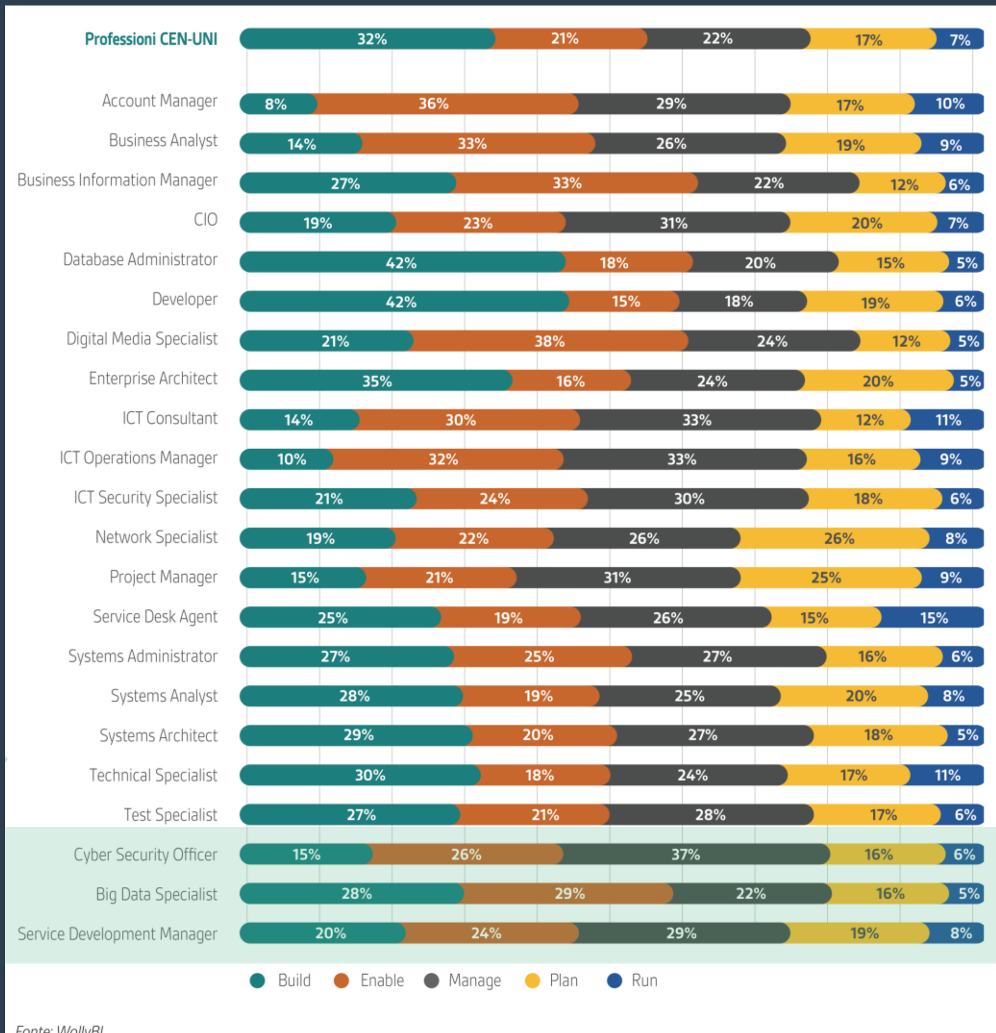
Variation 2019 vs 2018: +31%
Variation 2019 vs 2017: +149%

SOCIAL MEDIA SPECIALIST – 0,4k vacancies 2014-2019



Variation 2019 vs 2018: +105%

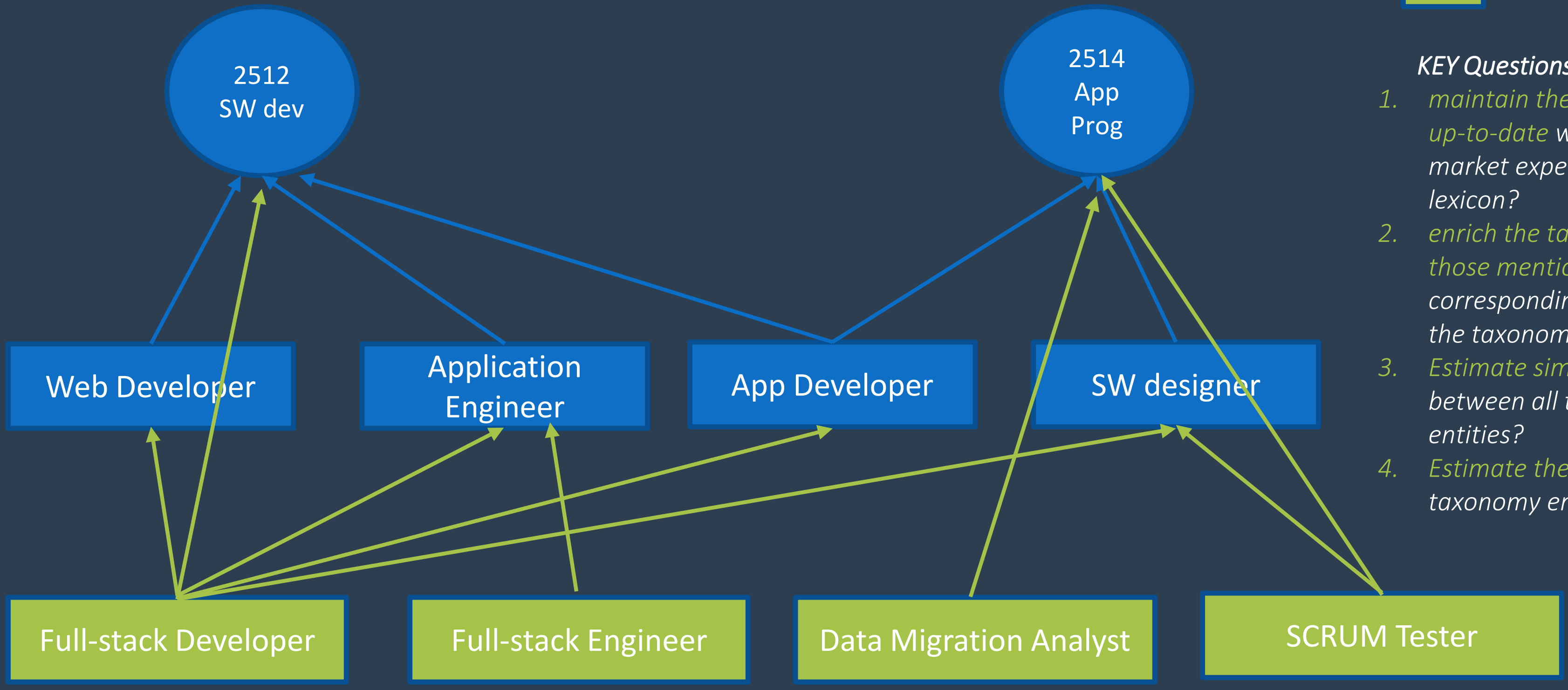
Variation 2019 vs 2017: +123%



New Occupations can be compared against traditional ones using different taxonomies
(the eCF Competence Framework in such a case)

Taxonomy Extension: How to improve skills/occupations taxonomies through semantic similarities within OJVs?

Main Idea



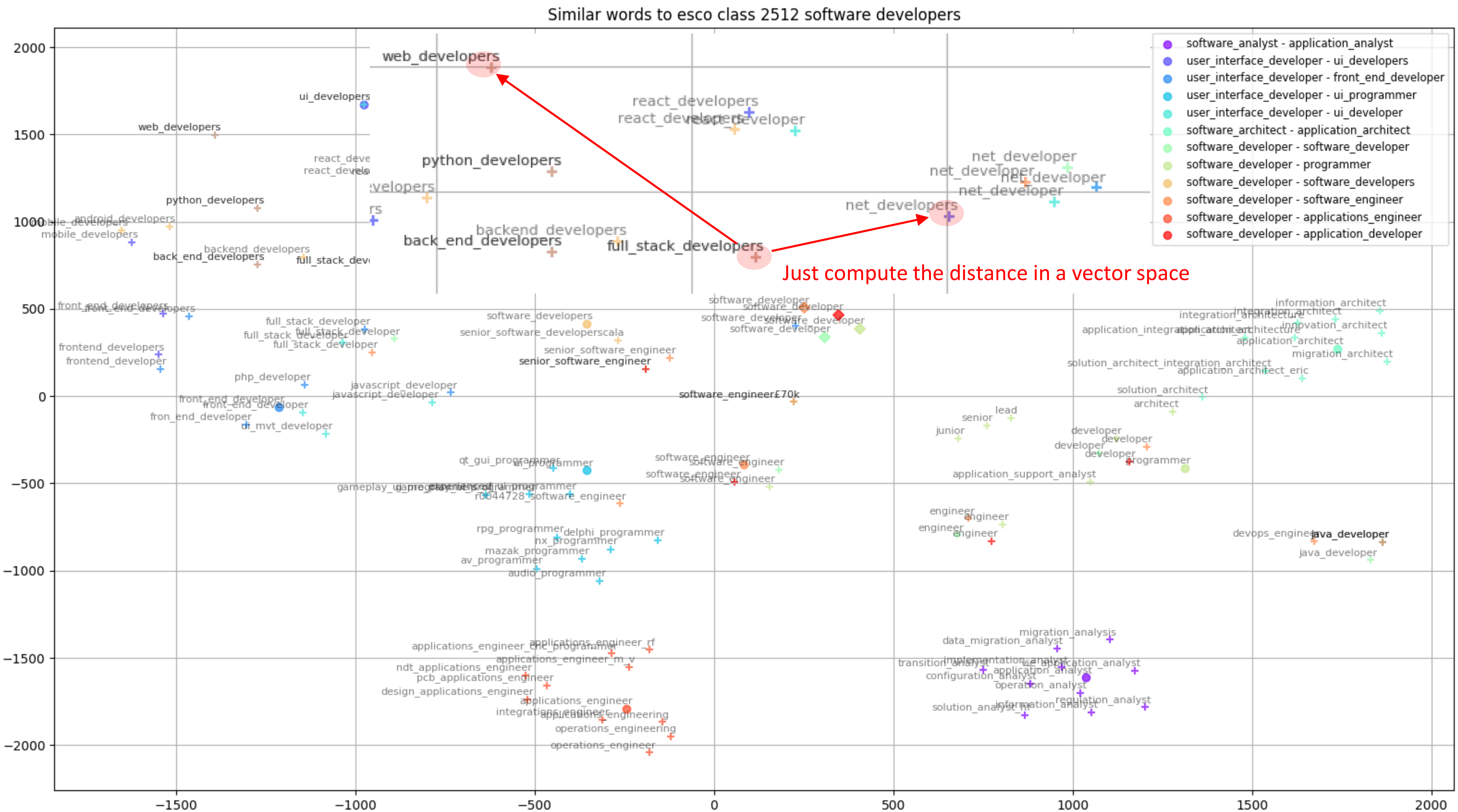
ISCO iv digit

ESCO v digit & alternative labels

Mentions from OJV

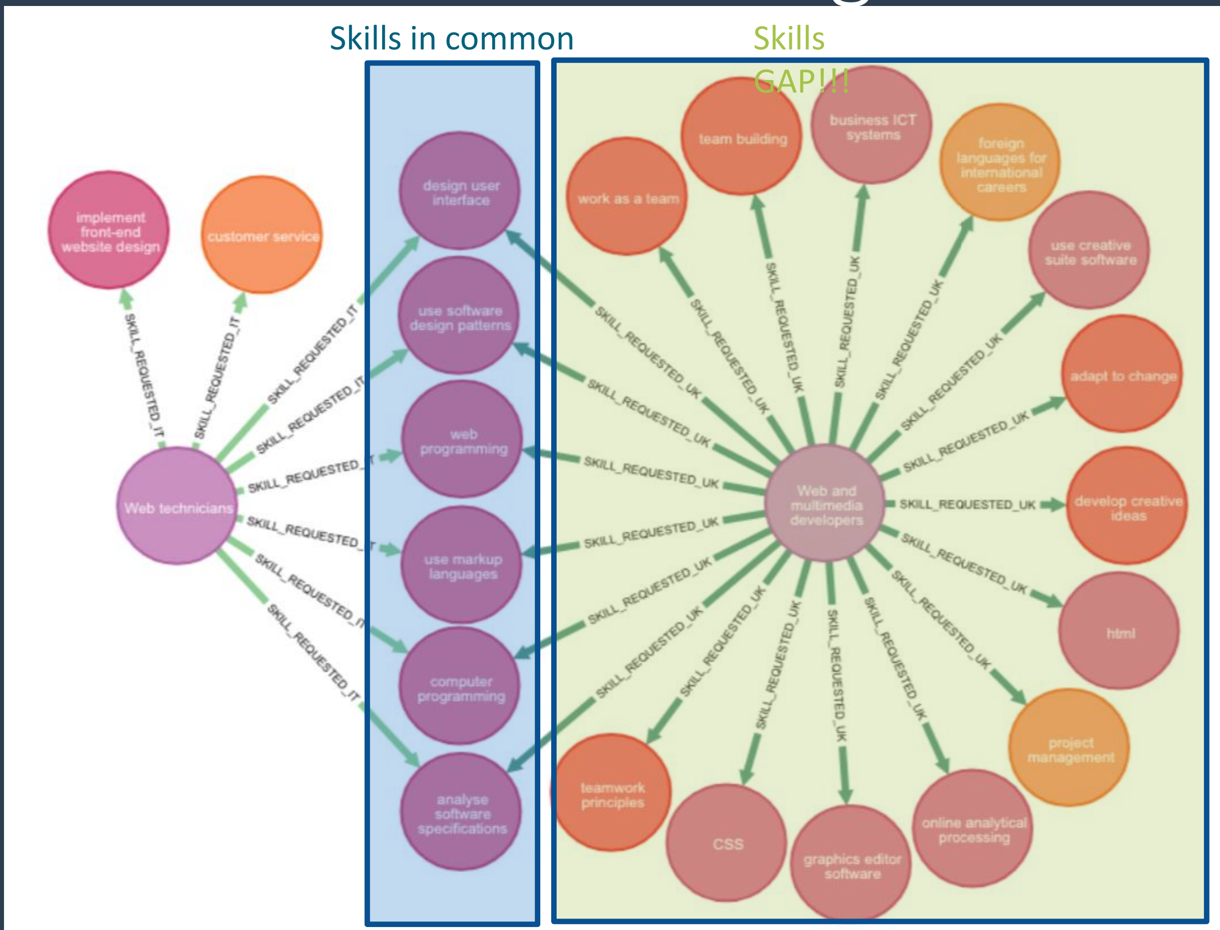
KEY Questions: How to…

1. *maintain the taxonomy up-to-date with labour market expectation and lexicon?*
2. *enrich the taxonomy with those mentions to corresponding entities in the taxonomy?*
3. *Estimate similarities between all taxonomy entities?*
4. *Estimate the relevance of taxonomy entities?*



Similar words to esco class 2512 software developers
software_analyst - application_analyst
user_interface_developer - ui_developers
user_interface_developer - front_end_developer
user_interface_developer - ui_programmer
user_interface_developer - ui_developer
software_architect - application_architect
software_developer - software_developer
software_developer - programmer
software_developer - software_developers
software_developer - software_engineer
software_developer - applications_engineer
software_developer - application_developer
web_developers
python_developers
back_end_developers
full_stack_developers
net_developers
react_developers
Just compute the distance in a vector space

Extending Skills Taxonomy



Compare different Web labour markets [IT, UK and DE here] to perform Skill Gap Analysis using the taxonomy as a baseline.

Question: *“starting from a given occupation of the ISCO taxonomy for the ITA labour market, what is the occupation in the UK labour market whose requested skills with a better fit?“*

Skills associated to “Web Technician” in ITA are more similar to a “Web and Multimedia Developer” in UK

Build-up an LMI Ontology

VIDEOS

Compare Different Countries (Web Technician)

OJV collected from UK, IT and DE in 2018

Nazione	Top 10 skill	*tf-idf* IT	UK	DE
IT	web programming	**100.0**	38.0	100.0
IT	analyse software specifications	**95.18**	36.84	21.72
IT	computer programming	**89.17**	48.82	59.35
UK	Marketing	39.69	**100.0**	56.56
UK	online analytical processing	0	**94.39**	0
UK	adapt to change	69.21	**87.21**	79.47
DE	web programming	100.0	38.0	**100.0**
DE	use markup languages	80.47	34.03	**82.97**
DE	adapt to change	69.21	87.21	**79.47**

Improve classification through visual explanations: eXplain and Validate

Explainable Labour Market Intelligence (XLMI)

Goal. **To improve the believability of the analysis and results provided to the final users by explaining the behaviour of AI algorithms used to produce them.**

Idea. ML algorithms act like a black-box, and there is no way to guess the reason behind a decision. eXplainable AI (XAI - launched by DARPA in 2016) aims at building a new generation of ML systems able to explain their decision in a human manner

Benefits. Improved ability in understanding and monitoring the classification process (even in case this is a complex pipeline of algorithms) and in identifying misclassification to improve both the accuracy and the system understanding.

Why do we need explanations?

No way to guarantee (or at least to make evidence) to users that a system learned the *“right model”*, but just it learned the *“model right”*

Explain (**in a human-readable format)** why a system classified an item on a class would make it more reliable and trustable to the final user

Be able to understand and isolate what forced a system to predict an outcome can be used in a human-in-the-loop validation process to improve:

- **Transparency** to the end users, that have to make decisions on the basis of your outcomes;
- **Believability and utility** of the deployed system;
- **Accuracy** of classification;
- And more (see after)

RA2: Improve classification through visual explanations: eXplain and Validate

Proposal. Use OJV data to text XLMI algorithms

Potential Stakeholders. National associations of Information Technology companies

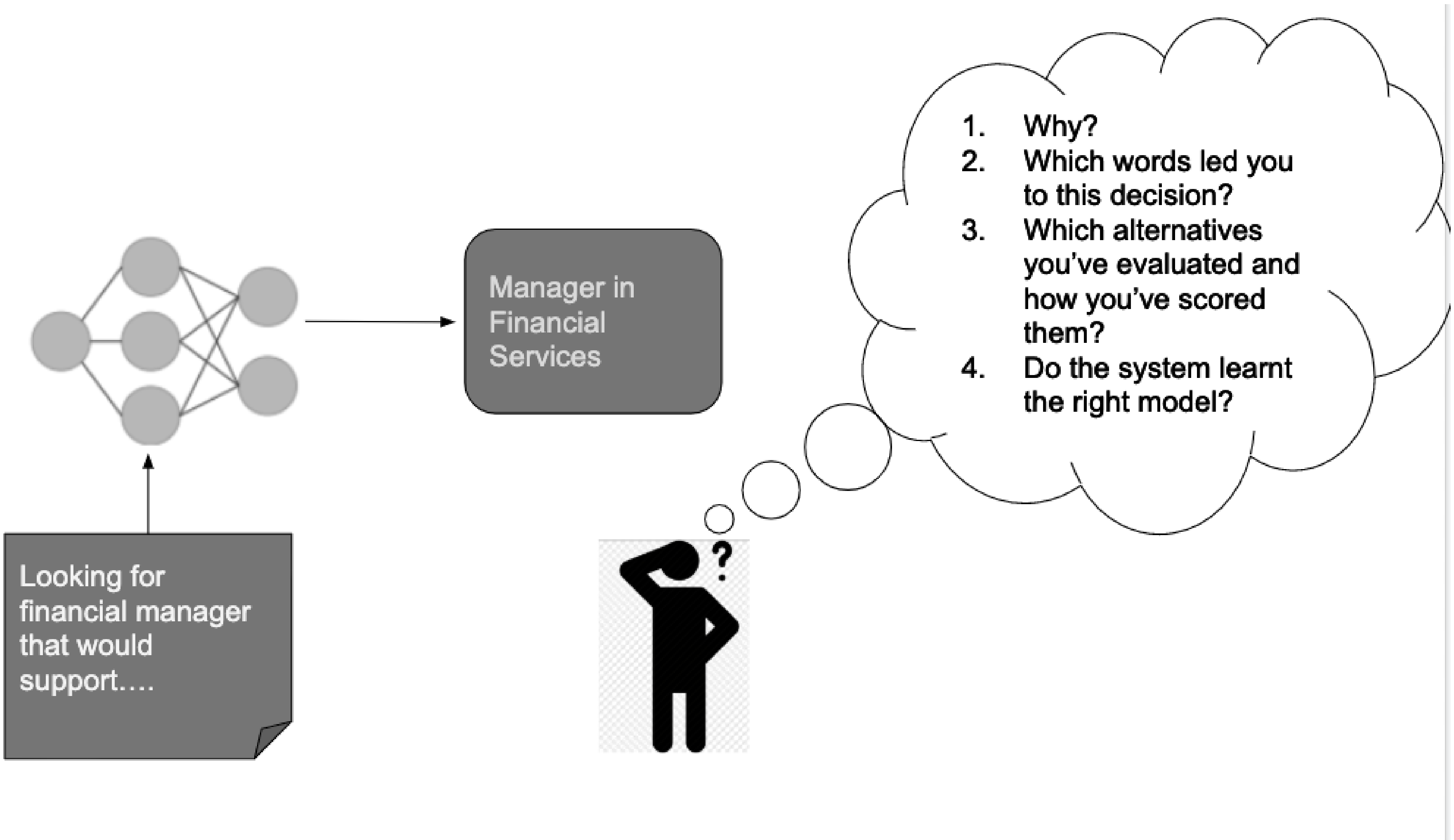
Research Phases

- Define a global/local model of Explainability
- Evaluate XAI algorithms
- Implement the XLMI algorithm
- Text the defined XLMI algorithm on the OJV data

Research Output.

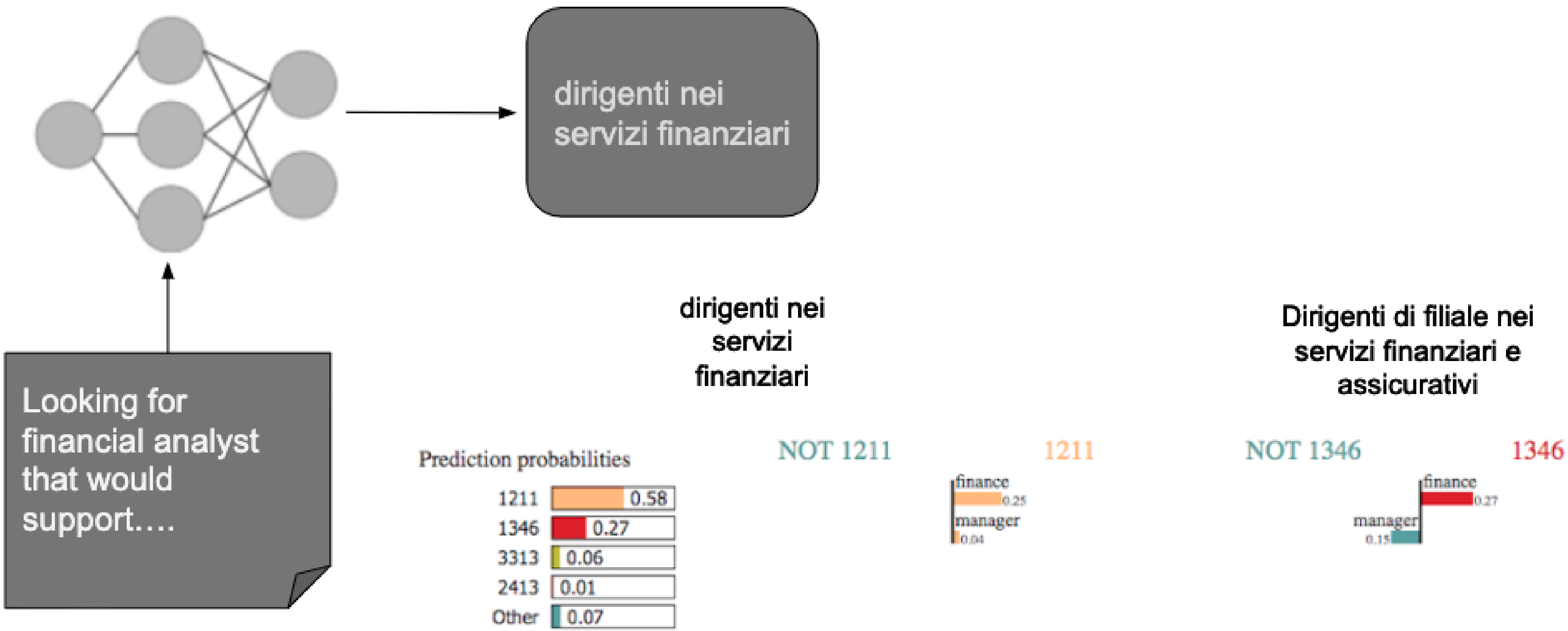
- Working Papers;
- A research proposal to explain WolliBI’s classification features [working].

Focus: eXplain and Validate

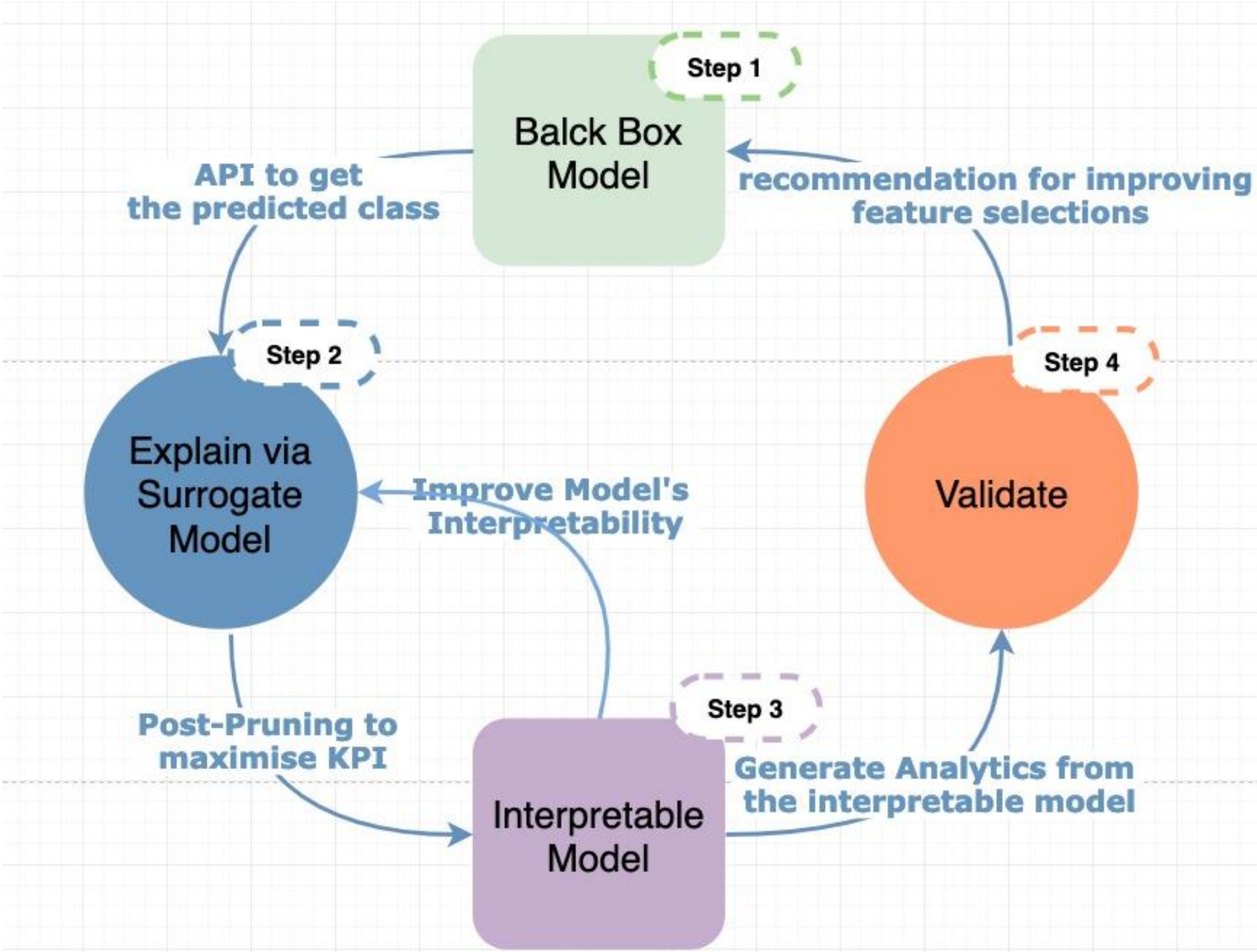


Focus: eXplain and Validate

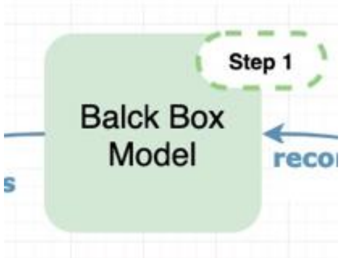
Might this help?



Focus: eXplain and Validate



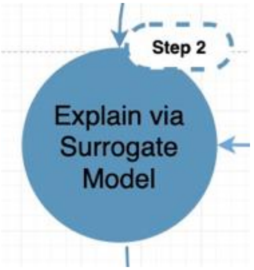
Focus: eXplain and Validate



Classify each job vacancy according to the system’s classification criteria

Notice: the preprocessing pipeline is unique modulo the language, this means some terms that do not account for the classification task (eg, senior, milan, etc…) are not discarded as they are useful to derive sectors, experiences, etc.

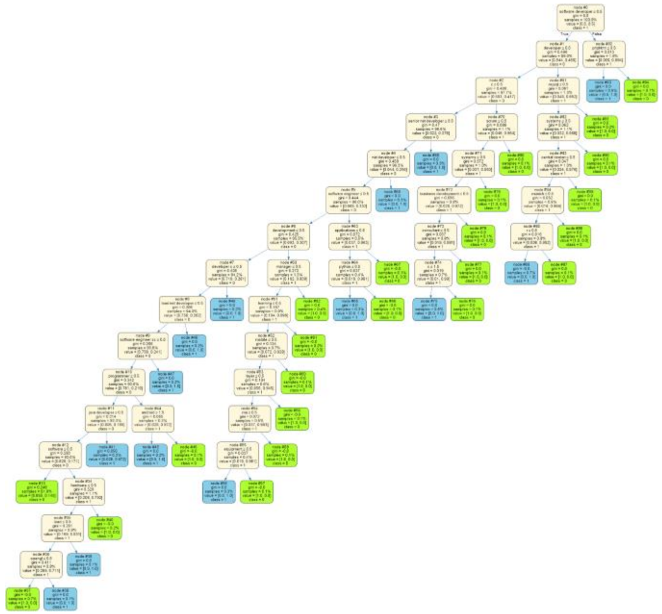
Focus: eXplain and Validate

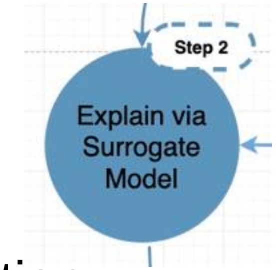


Use global interpretable models to explain each class (ie, one of 400+ ESCO occupations)

The result is a tree that shows the system’s behaviours in classifying vacancies belonging to that ESCO occupation

Software Developer →



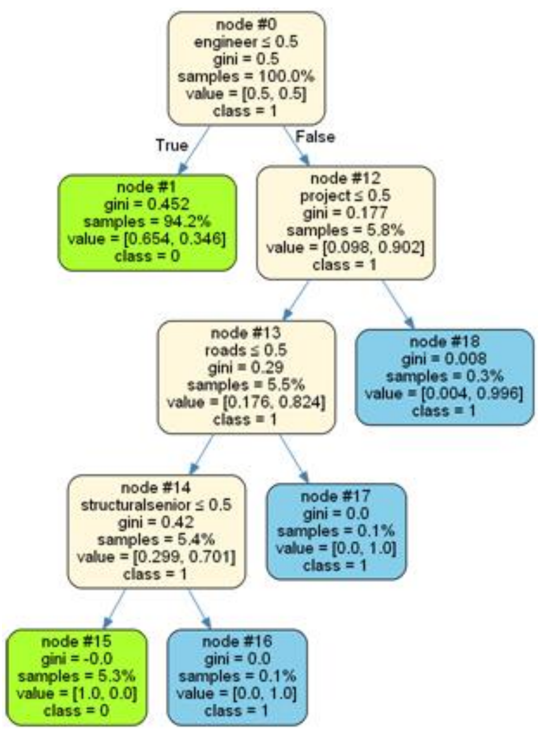


Focus: eXplain and Validate

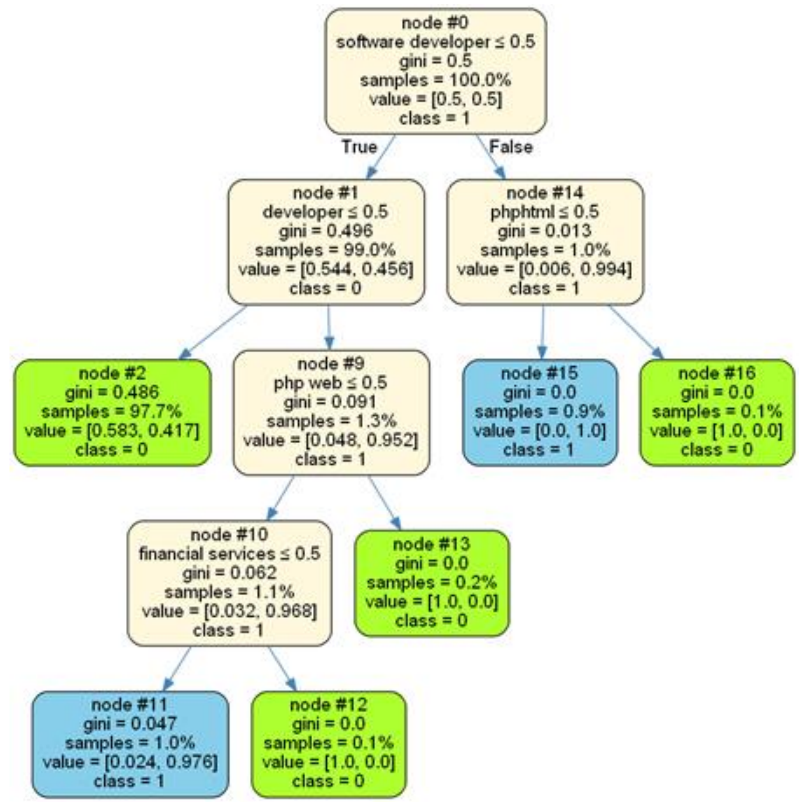
Basically, this is a way to see under the hood of the classification system

It can be seen as a set of nested if-then-else that reveal the system criteria

Civil Engineer



SW developer



Focus: eXplain and Validate

Whi

(

(

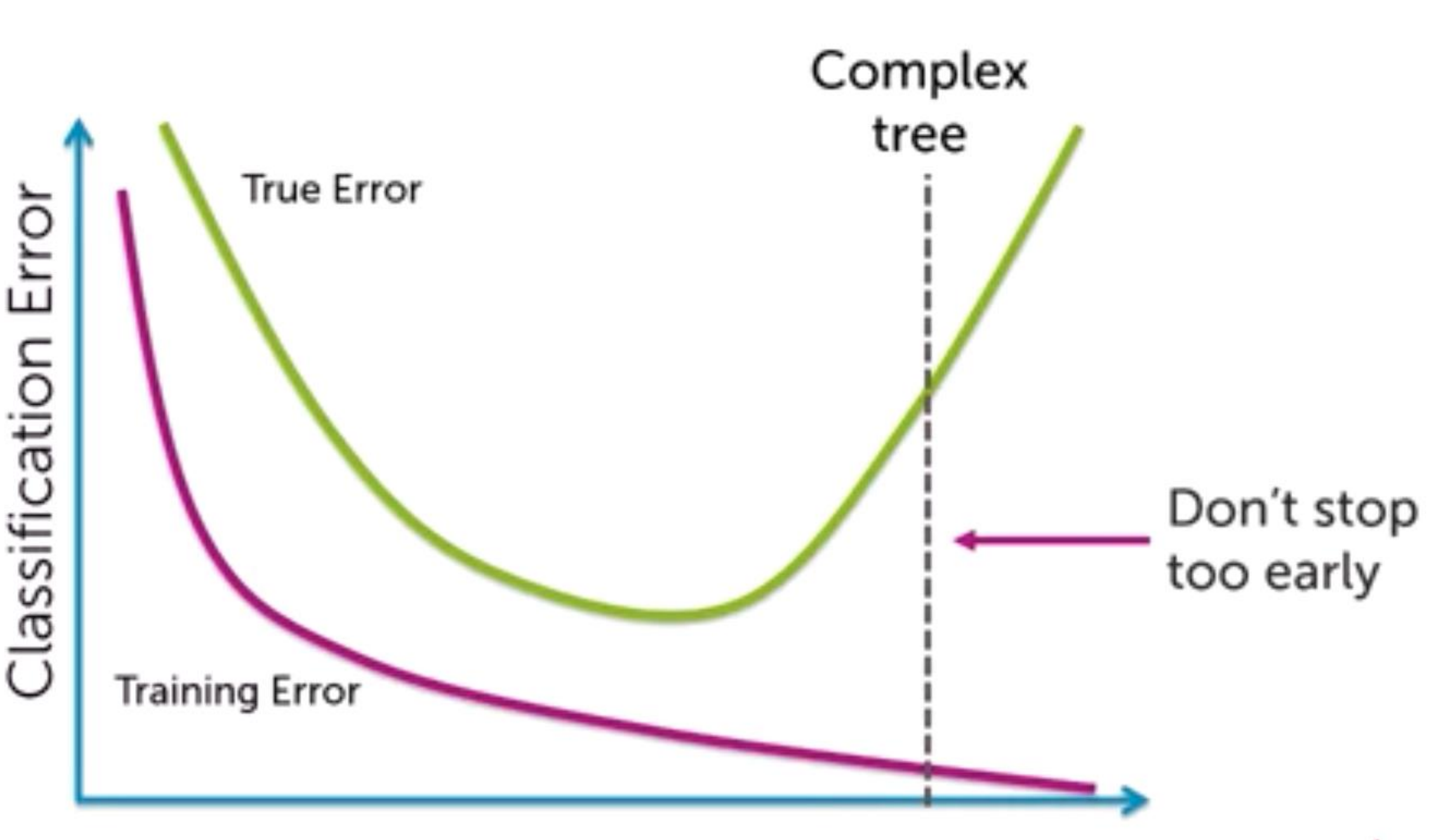
IDE

acc y)

and

At t

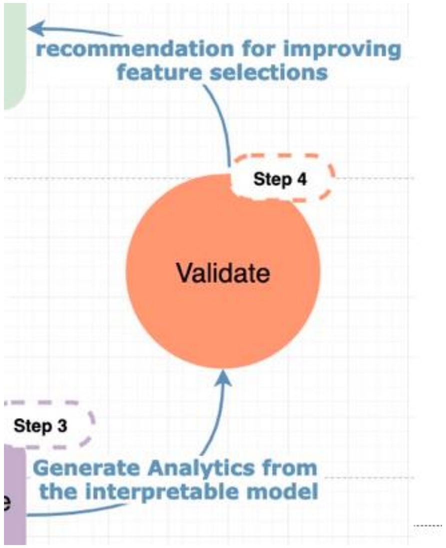
by t



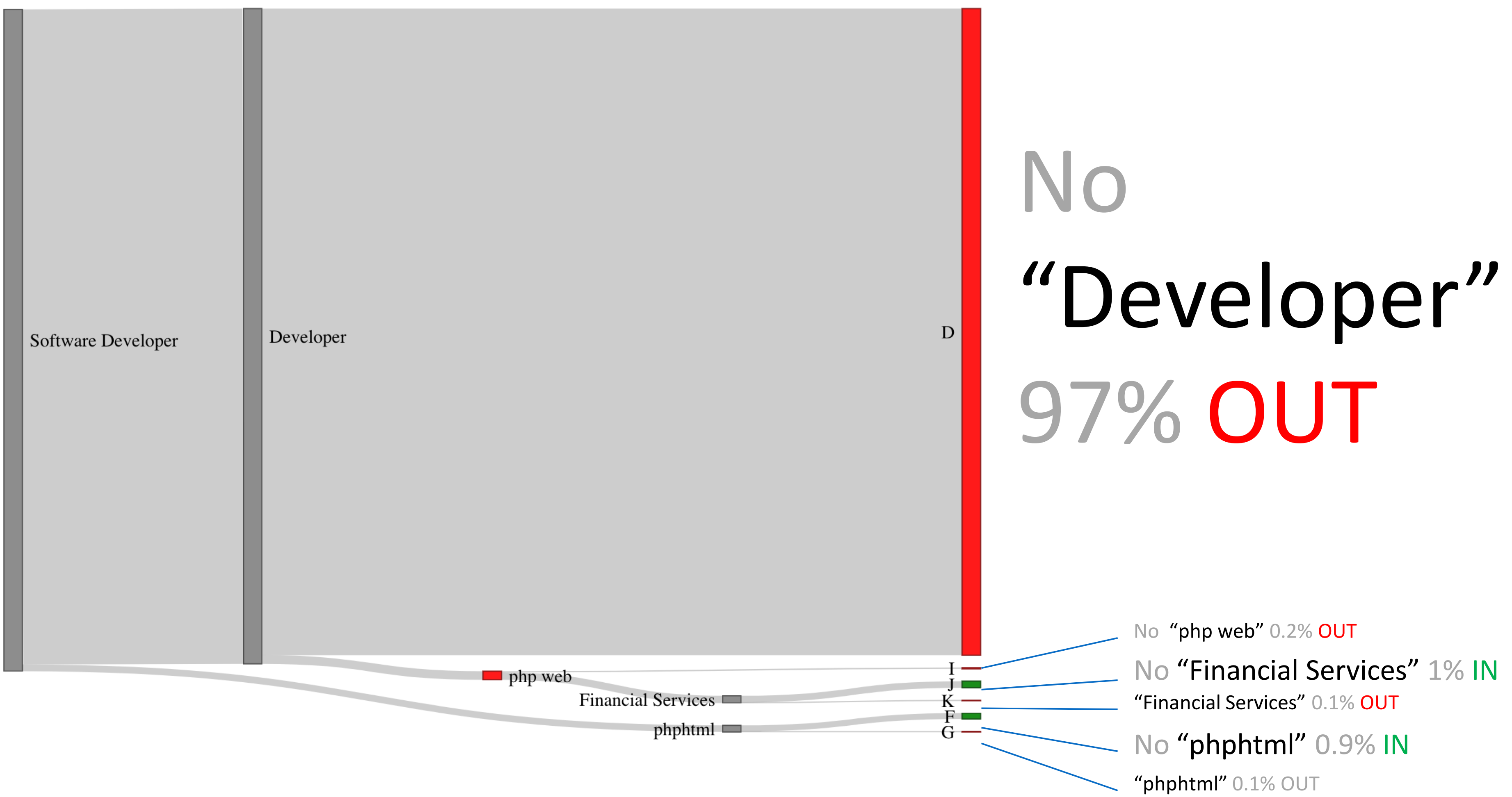
Focus: eXplain and Validate

Moving from **interpretations** to **explanations**

Once the system has been interpreted, we use visualisations to derive explanations in a human-readable manner



Focus: eXplain and Validate



Challenge

How to deal with representativity issue? [live]