Big Data for Labour Market Information

Session 4
Data cleaning: purposes and tools. Importance for quality and believability of the data and analysis.

Alessandro Vaccarino – Fabio Mercorio

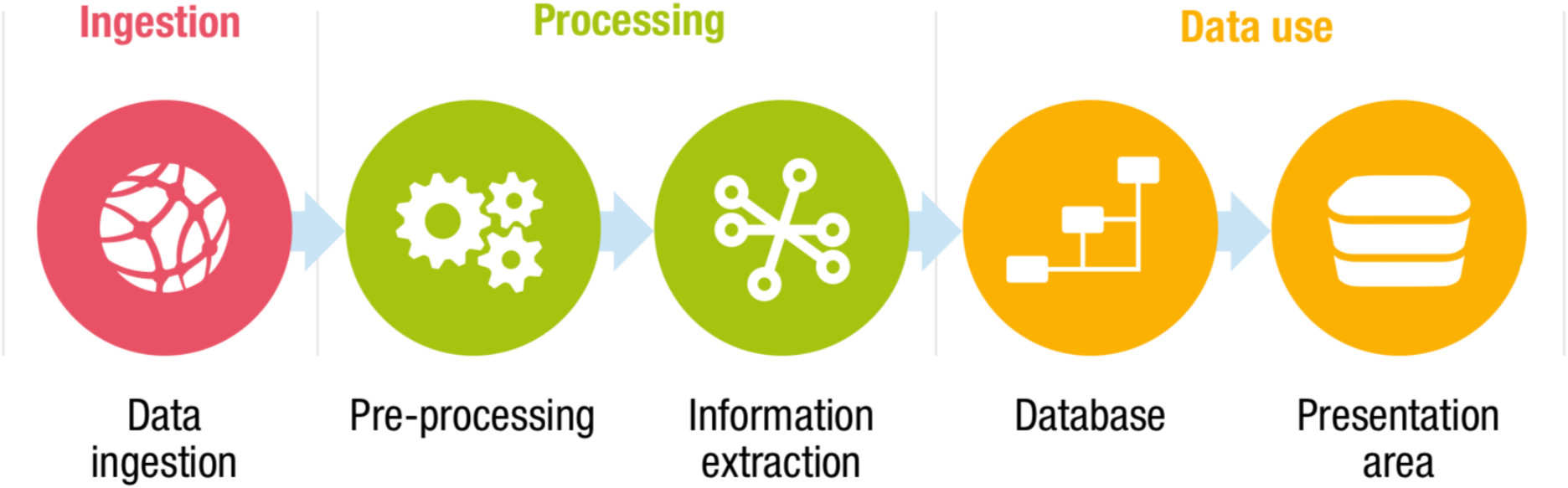
Big Data for Labour Market Information – focus on data from online job vacancies – training workshop
Milan, 21-22 November 2019

Topics

1. Challenges
    1. Data ingestion techniques
    2. Data processing pipeline
    3. Classification techniques

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


Ingestion
Processing
Data use
Data ingestion
Pre-processing
Information extraction
Database
Presentation area

Topics

1. Challenges
    1. Data ingestion techniques
    2. Data processing pipeline
    3. Classification techniques

Data Ingestion phase

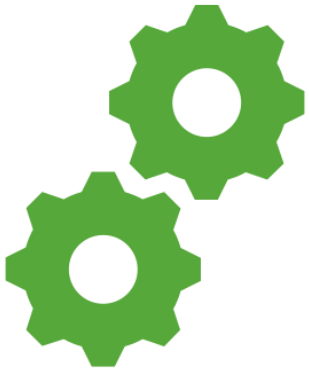
The process of obtaining and importing data from web portals and storing them in a Database

Focus on volumes

Coverage augmentation & maximization

Direct agreements with the most relevant sources

Ingestion Challenges

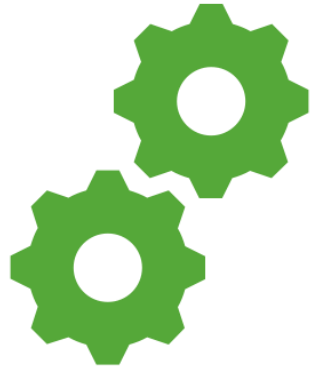


Robustness of the process

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

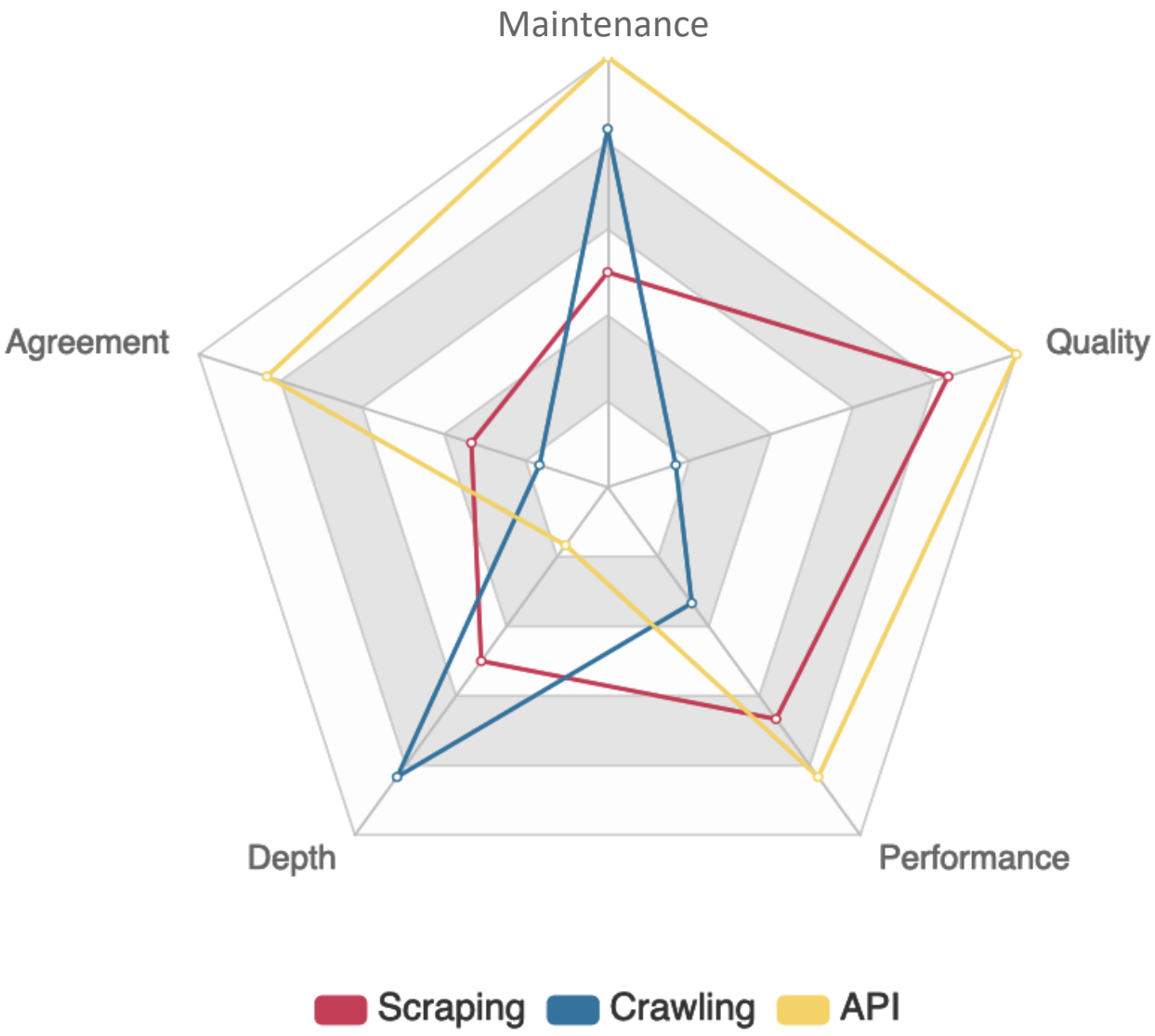
Risk: loss of quality

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

- API
- Scraping
- Crawling

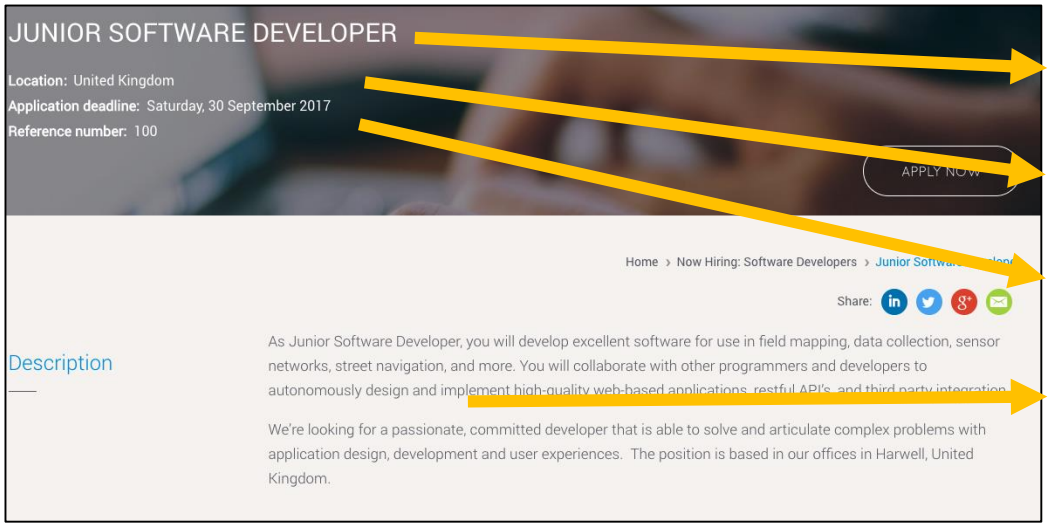
Ingestion Challenges - Quality

- 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


Maintenance
Quality
Performance
Depth
Agreement
Scraping
Crawling
API

Scraping – An example

Web scraping is data scraping used for extracting structured data from websites



Title: Junior Software Developer

Area: United Kingdom

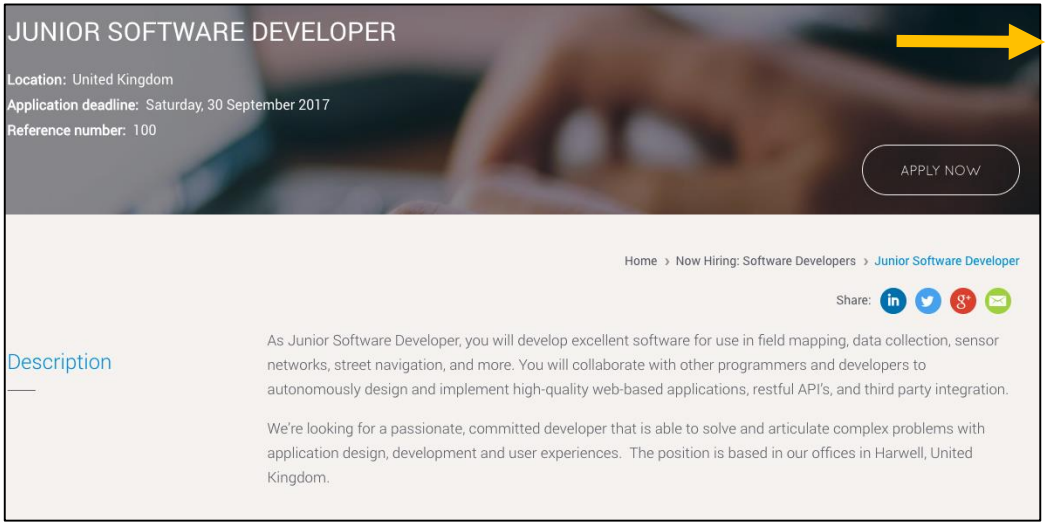
Time: Saturday, 30 September 2017

Description: As Junior Software Developer, you will develop excellent software for use …

Crawling – An example

A Web crawler is a bot that systematically browses web portals for the purpose of download all their pages.

Crawling is the most common way to get information massively from the Internet: search engine spiders (e.g. GoogleBot)



Web page:

```
<!DOCTYPE html>
  <head>
    <meta name="title" content="Junior
Software Developer" />
  </head>
  <body>
    <header>
      <h2>Junior Software Developer</h2>
      <div><div>Location</div>United
Kingdom</div>
    …
    </header>
    <div><div>Description</div>
    <span>As Junior Software Developer, you
will develop excellent software for use…
```

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

Ingestion Challenges - Scaling

We developed a solution based on microservices, that creates and deletes "virtual browsing computers" as needed. Each computer has multiple browsers that can emulate human web navigation.

Main differences with a real computer are:

1. They don't have a monitor, but saves pages on our Data Lake
2. We can scale up and down as needed

Topics

1. Challenges
    1. Data ingestion techniques
    2. Data processing pipeline
    3. Classification techniques

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

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⟩.

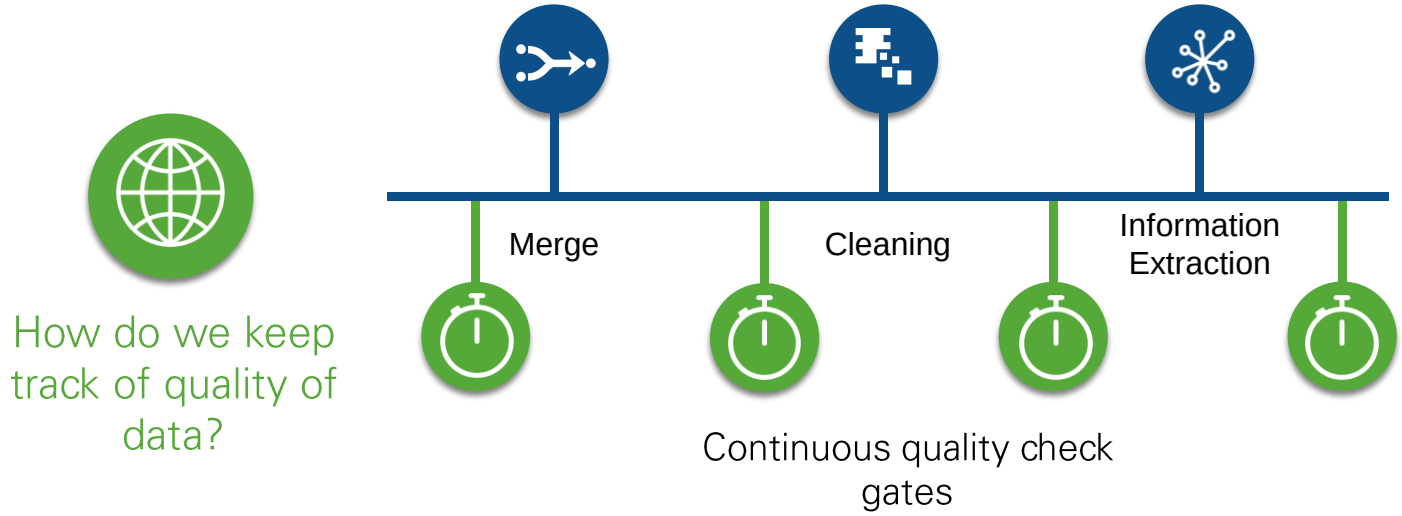
We're looking for a passionate, committed ⟨developer⟩ that is able to ⟨solve⟩ and articulate ⟨complex problems⟩ with ⟨application design⟩, ⟨development⟩ and ⟨user experiences⟩.
The position is based in our offices in ⟨Harwell⟩, ⟨United Kingdom⟩.

Data Pre-Processing
What to do with noise?

We don't physically delete noise

We collect it to keep track of the overall process, and monitor:

- Noise type → To identify need to develop some deeper quality check process
- Noise trends → To detect sources that are increasing/decreasing noise and deal it
- Analytical purposes → Analyse country-specific cultural environments, like the use of OJVs portal to promote training courses
- Monitoring → Keep track of the overall process



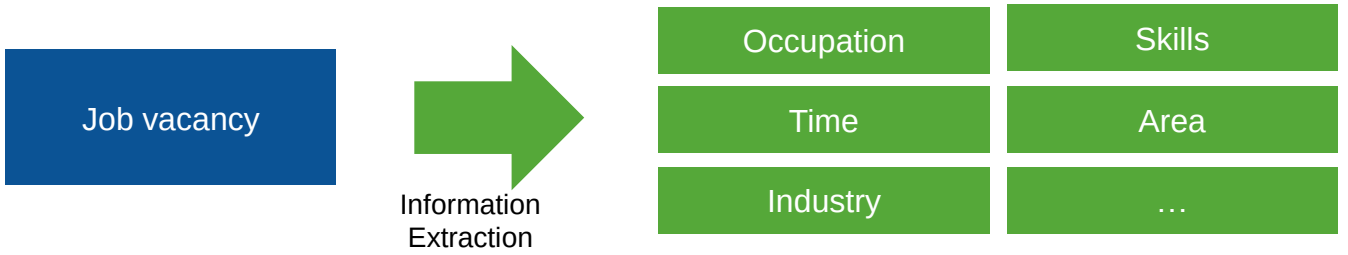
Topics

1. Challenges
    1. Data ingestion techniques
    2. Data processing pipeline
    3. 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.

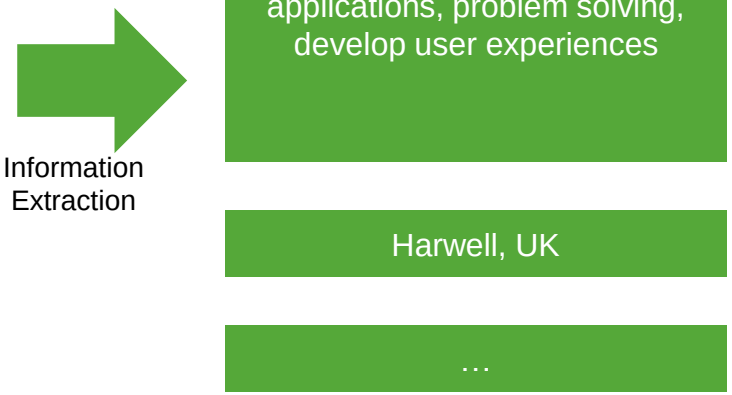
Data Classification - An example



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.

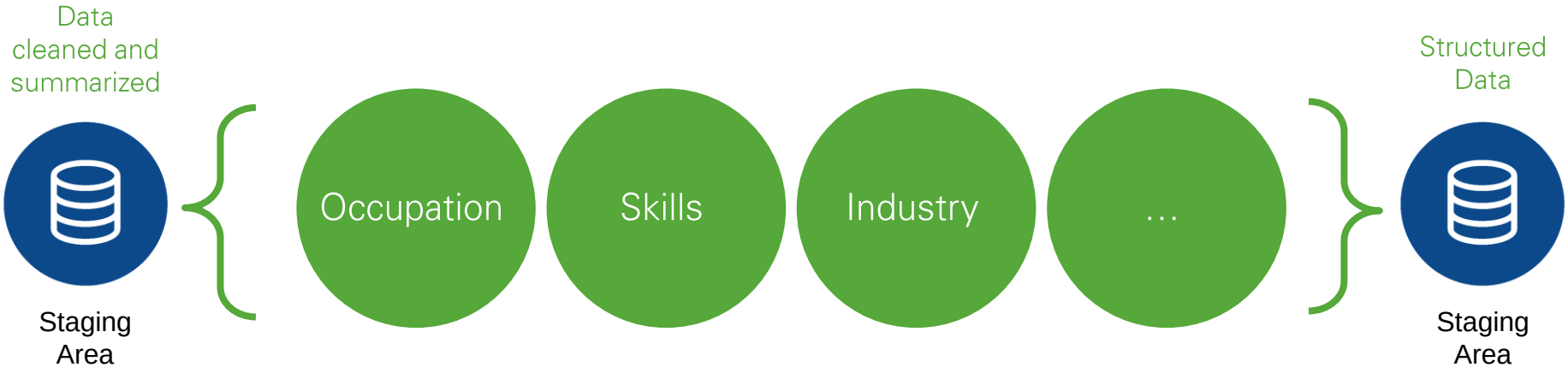
We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.



Information Extraction and Classification
Real Time Labour Market Intelligence

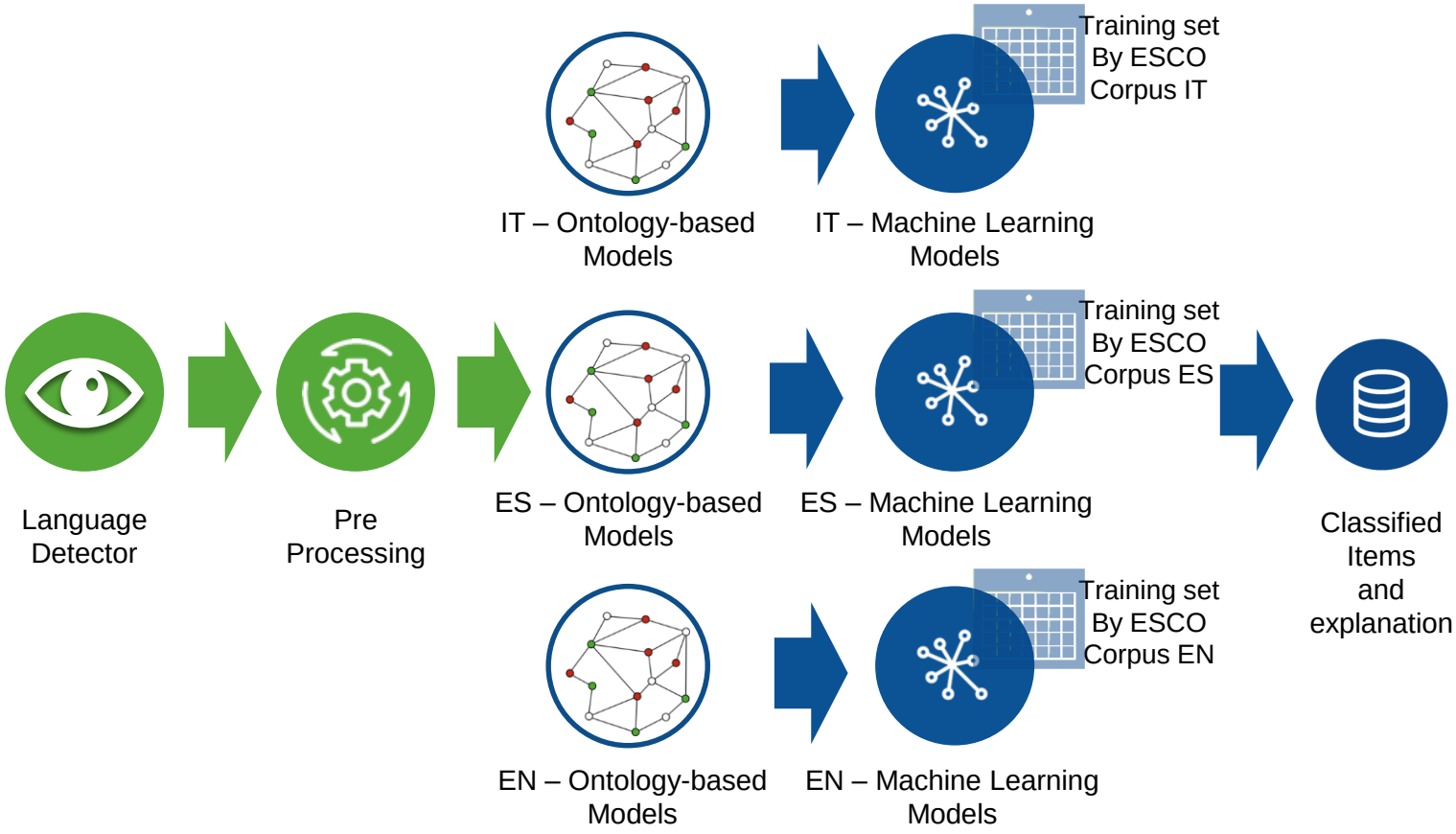
Information Extraction is an area of natural language processing that deals with finding factual information in free text.

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



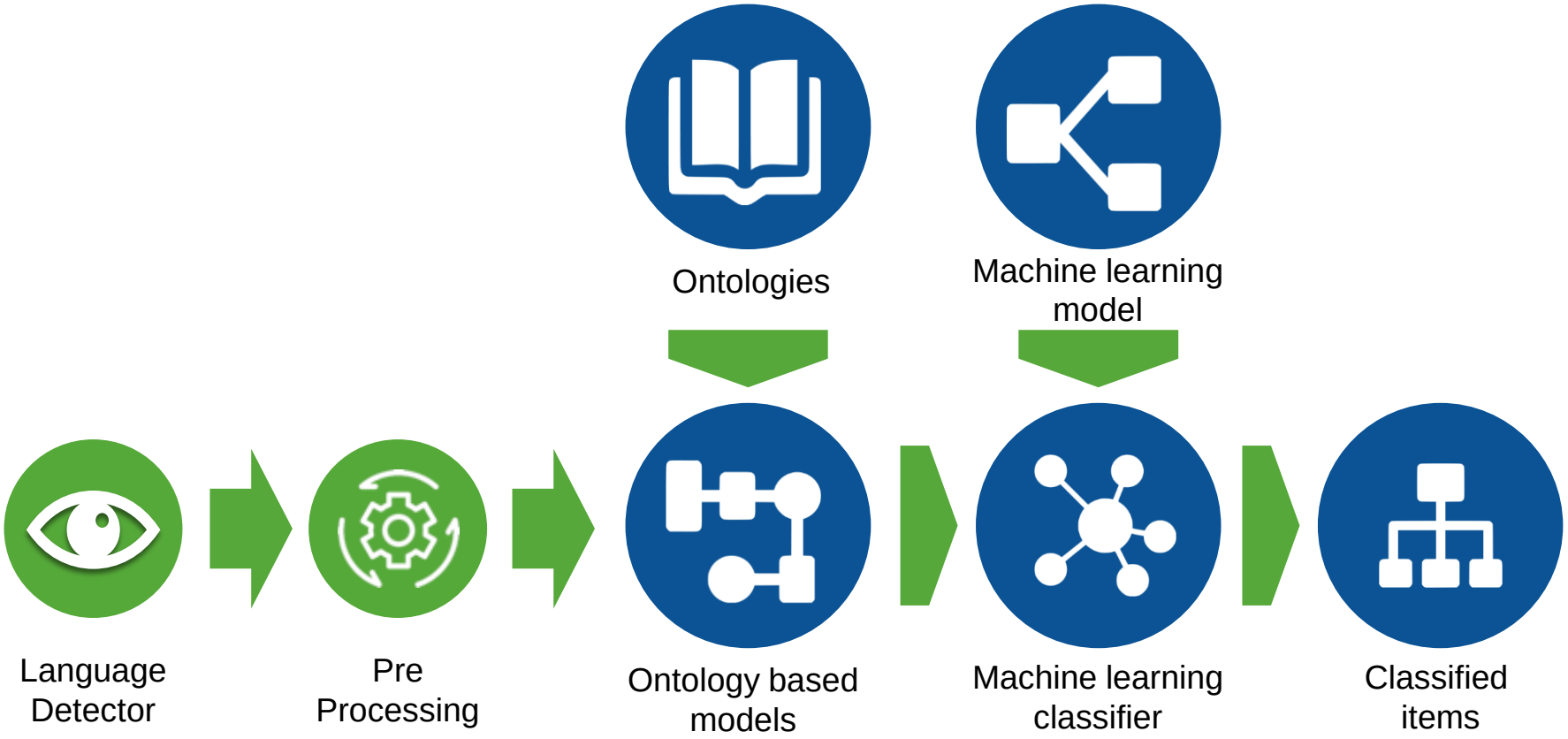
Machine Learning → Ontology based learning, supervised learning and unsupervised learning, etc.

Classification



What does “Ontology-based Models” means?
How we can use ontologies to classify?

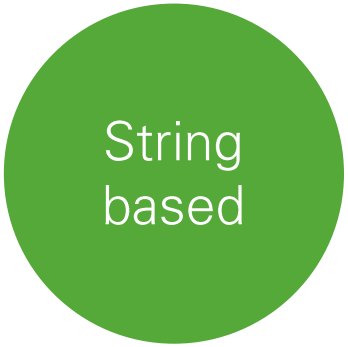
Occupations pipeline



Considerations on Occupation Classifier

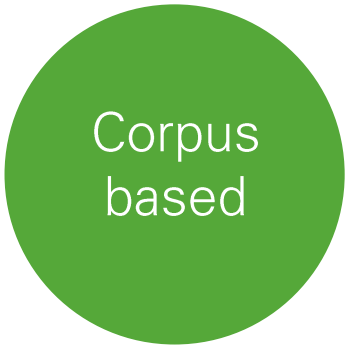
- Ontology based learning + Supervised learning
  - Esco Ontology
  - New labels from Topic modelling
- One model for each language
- Data labelled by expert from each country
  - ~100k job ads (cleaned train set using our ontology)
  - 436 possible targets
- Evaluating set 20% of gold dataset job ads
  - Weighted Precision ~86%
  - ~430 detected professions

Text Similarity Approaches



String similarity measures operate on string sequences and character composition.

Jaro-Winkler, Jaccard, Cosine similarity



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

Sector

Job Reference: 990-NHSE8576N
Industry: Health
Salary: 56,665 – 69,168 per annum
Location: Leeds

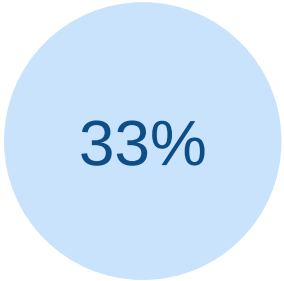
NHS England leads the National Health Service (NHS) in England. We set the priorities and direction of the NHS and encourage and inform the national debate to improve health and care. We want everyone to have greater control of their health and their wellbeing, and to be supported to live longer, healthier lives by high quality health and care services that are compassionate, inclusive and constantly-improving…

Nace was present as structured field in OJV (valued as "Health", that matches ontology)

Classification value:
86 - Human health activities

Reference

Structured data



Record linkage with NACE

Contract

Labourer - Aylesbury
Contract: Temporary (3 Month)
Salary: £10 per hour

We are currently looking for a hard working an honest labourer in the Aylesbury area.
You will be the main site labourer with duties including cleaning the site and helping trades out around the site. You can access this site with public transport, and has parking access if you….

Contract was present as structured field in OJV (valued as "temporary", that matches ontology value)

Classification value:
Temporary

Reference

Custom taxonomy

Permanent
Self Employment
Temporary

Structured data

30%

Record linkage with taxonomy

76%

Record not linked (unclassified)

24%

Working Hours

Job Reference: 184-SS.GEN.38
Department: Dementia
Location:Bracken House, Chard

The Chard Older Persons Community Mental Health Team are actively seeking **to recruit a Part time** Band 5 Community Mental Health Nurse to assist with the Memory Assessment Service and Day Hospital.
As part of an innovative Integrated Team you will be working closely with District Nursing, Integrated Rehab team and the Medical team as well as GP's, Adult Social care, Acute sector and Voluntary sector….

Working Hours was not present as structured field in OJV, but text contains reference to working hours ("part time") that matched ontology

Classification value:
Part Time

Reference

Full time
Part time

Custom taxonomy

Structured data

29%

Record linkage with taxonomy

63%

Record not linked (unclassified)

37%

Educational Level

Role: Rolling Stock Team Leader
Location: South London
Salary: Approx. £47,500
Education requirements: Associate Degree
Experience: Less than 1 Year

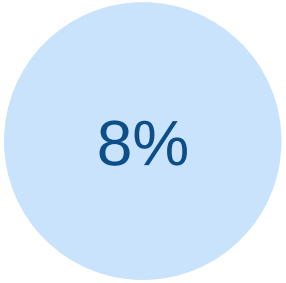
The purpose, to lead the day to day activities to achieve timely stock delivery whilst ensuring that both relevant maintenance standards are achieved and passenger environment activities are enabled. Roles and responsibilities include but are not limited to: Daily delivery of the fleet into service, reliably and consistently To be part of the leadership team that delivers a cost effective and efficient maintenance …

Educational Level was present as structured field in OJV as "Education Requirements" (valued as "Associate degree", that matches ontology's alternate title)

Classification value:
Bachelor or equivalent

Reference

Structured data



Record linkage with ISCED 2011

Salary

Role: Rolling Stock Team Leader
Location: South London
Salary: Approx. £47,500
Education requirements: Associate Degree
Experience: Less than 1 Year

The purpose, to lead the day to day activities to achieve timely stock delivery whilst ensuring that both relevant maintenance standards are achieved and passenger environment activities are enabled. Roles and responsibilities include but are not limited to: Daily delivery of the fleet into service, reliably and consistently To be part of the leadership team that delivers a cost effective and efficient maintenance …

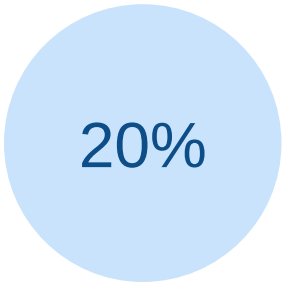
Salary was present as structured field in OJV (valued as "£47,500 per Year" and converted to EUR currency)

Classification value:
48.000 - 54.000 EUR Per Year

Reference

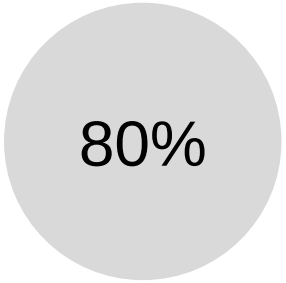
13 levels

Structured data



Record linkage with taxonomy

Record not linked (unclassified)



Experience

Role: Rolling Stock Team Leader
Location: South London
Salary: Approx. £47,500
Education requirements: Associate Degree
Experience: Less than 1 Year

The purpose, to lead the day to day activities to achieve timely stock delivery whilst ensuring that both relevant maintenance standards are achieved and passenger environment activities are enabled. Roles and responsibilities include but are not limited to: Daily delivery of the fleet into service, reliably and consistently To be part of the leadership team that delivers a cost effective and efficient maintenance …

Experience was present as structured field in OJV (valued as "Less than 1 Year", that matches ontology)

Classification value:
Up to 1 year

Reference

8 levels

Custom taxonomy

Structured data

5%

Record linkage with taxonomy

43%

Record not linked (unclassified)

57%

Place

Job Reference: 990-NHSE8576N
Industry: Health
Salary: 56,665 – 69,168 per annum
Location: Leeds

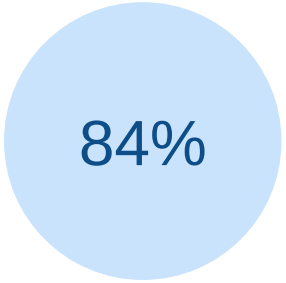
NHS England leads the National Health Service (NHS) in England. We set the priorities and direction of the NHS and encourage and inform the national debate to improve health and care. We want everyone to have greater control of their health and their wellbeing, and to be supported to live longer, healthier lives by high quality health and care services that are compassionate, inclusive and constantly-improving…

Place was present as structured field in OJV (valued as "Leeds", that matches ontology)

Classification value:
Leeds

Reference

Structured data



Record linkage with NUTS & LAU

Occupation

Unix Technician

In this role you will be responsible for these activities:
o Install and support the server operating system, system management software and operating system utilities
o Manage the operating system configuration
o Manage file systems and print queues
o Monitor and maintain operating system log files o Recommend operating system updates and configuration modification …

Machine Learning algorithm matched the correct Occupation, not present in ontology

Classification value:
2522 - Systems administrators

Reference
(ESCO 4th level)

ESCO v1
ISCO

Structured data

7%

Record linkage with ESCO/ISCO

100%

Skill

Are you an experienced Administrator, seeking your next contract in the Bristol area?
My client is a large property maintenance specialist with an immediate opportunity for a Branch Administrator to join the team on an initial interim basis.
The successful candidate will complete a range of **administration** tasks, including **answering incoming calls, liaising with contractors and raising invoices**.
Responsibilities:

- **Use the I.T systems** to provide an administration service in the preparation, processing and selection of estimates, bids and tenders
- **Ordering** of goods, materials and services to enable the requirements of contracts are met
- **Deal with internal and external communications** and record and or report information as necessary
- Ensure all necessary contract data, documentation and reports are accurate and produced on time
- Support Management in **meeting the business needs**.
- **Deal with Client / Customer queries and or communications** professionally and efficiently.

Requirements:

- Confident IT skills, **proficient in the use of MS Office**
- Excellent **communication skills** both written and verbal
- Must be an excellent organiser with proven **time management skills**
- ….