ETF Working together Learning for life

Big Data for Labour Market Intelligence

Day 1, Session 2

The Role of AI in the Data System

Alessandro Vaccarino – Mauro Pelucchi

22 November 2021

burningglass® TECHNOLOGIES

Topics

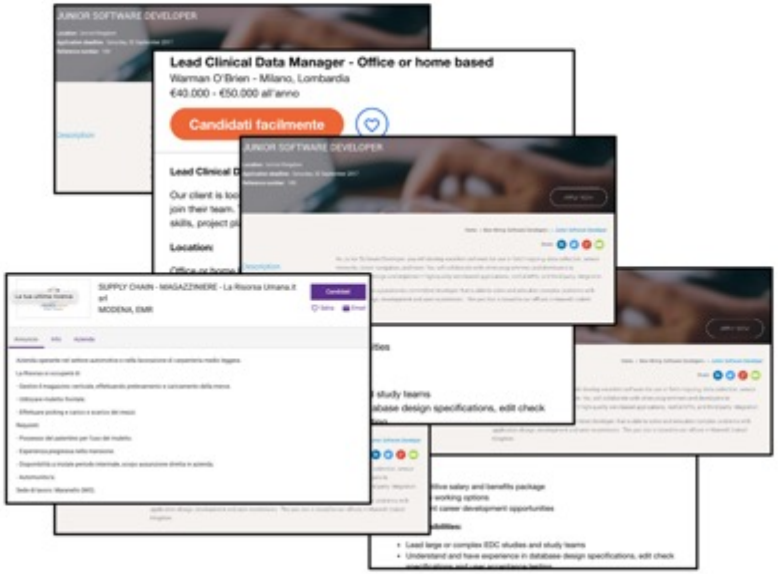
1. Recap
2. The Data System
    1. The functional architecture
    2. Data ingestion techniques
    3. Data processing pipeline
    4. Classification techniques

Topics

1. **Recap**
2. The Data System
    1. The functional architecture
    2. Data ingestion techniques
    3. Data processing pipeline
    4. Classification techniques

Our Goal

Transform Online Job Advertisments…



JUNIOR SOFTWARE DEVELOPER

Lead Clinical Data Manager - Office or home based

Warman O'Brien - Milano, Lombardia

€40.000 - €50.000 all'anno

Candidati facilmente

Lead Clinical D

Our client is loo
join their team.
skills, project pl

Location:

Office or home

JUNIOR SOFTWARE DEVELOPER

SUPPLY CHAIN - MAGAZZINIERE - La Risorsa Umana.it srl
MODENA, EMR

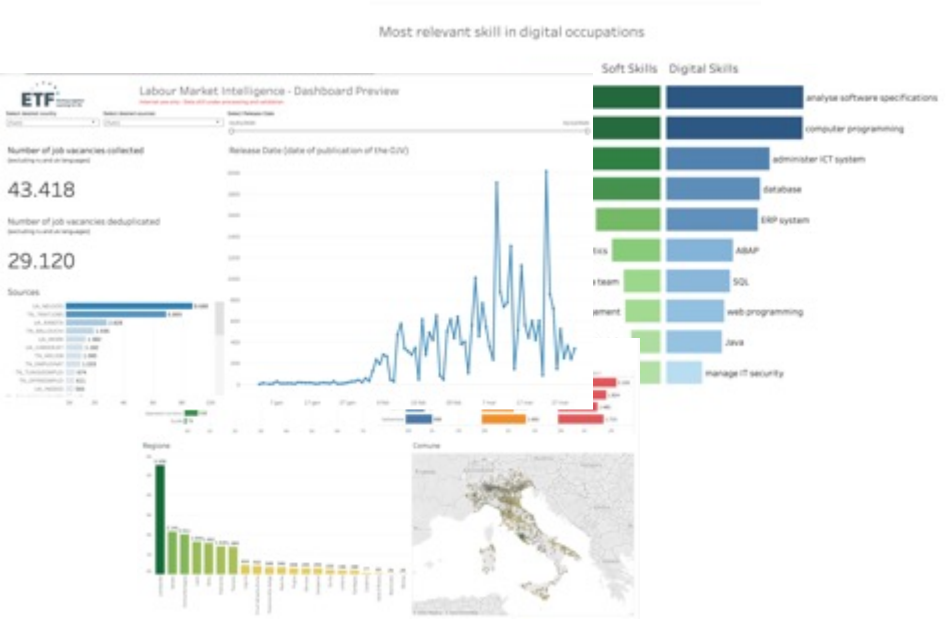
l study teams
abase design specifications, edit check

tive salary and benefits package
working options
t career development opportunities

abilities:

- Lead large or complex EDC studies and study teams
- Understand and have experience in database design specifications, edit check

…in insights and analytics

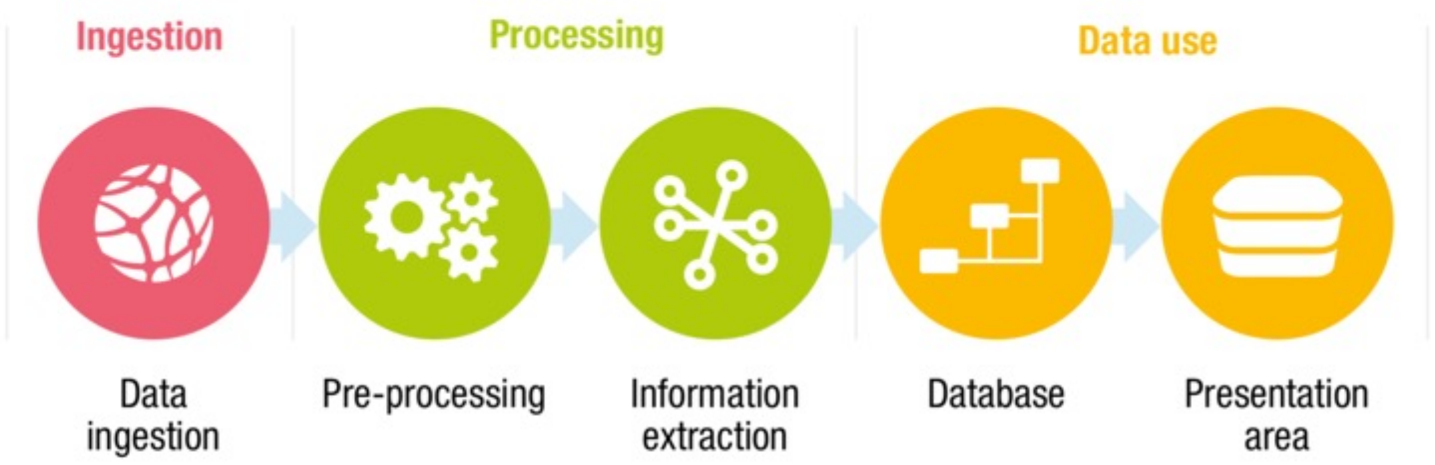


Challenges

- 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

Our Approach

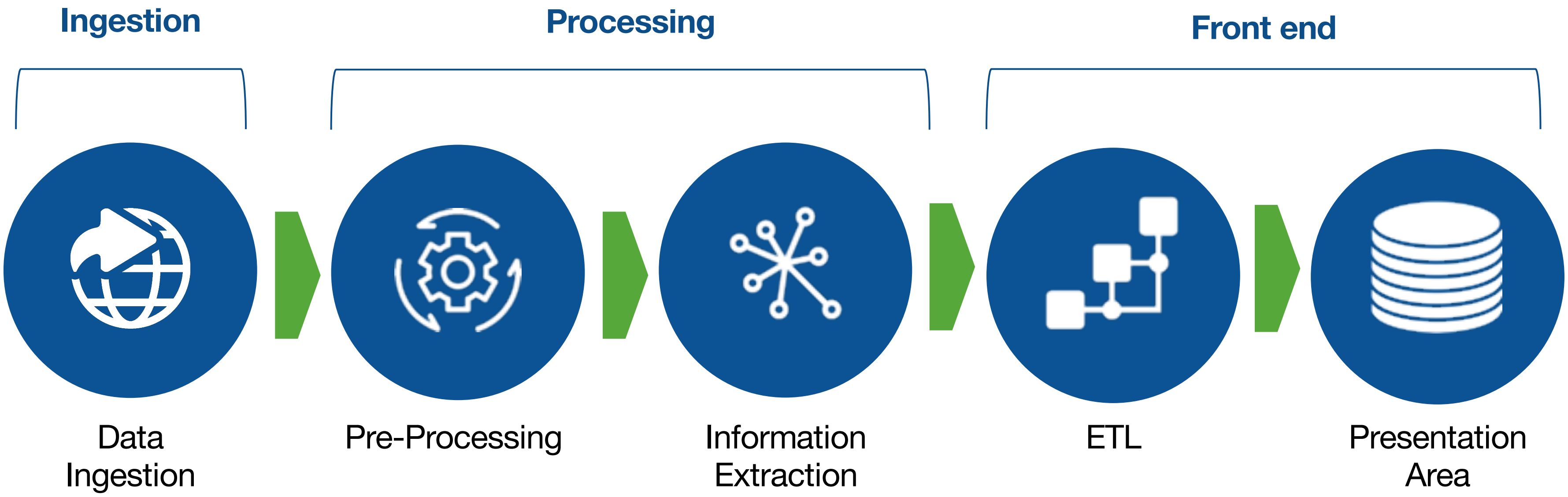
KDD 4 LMI



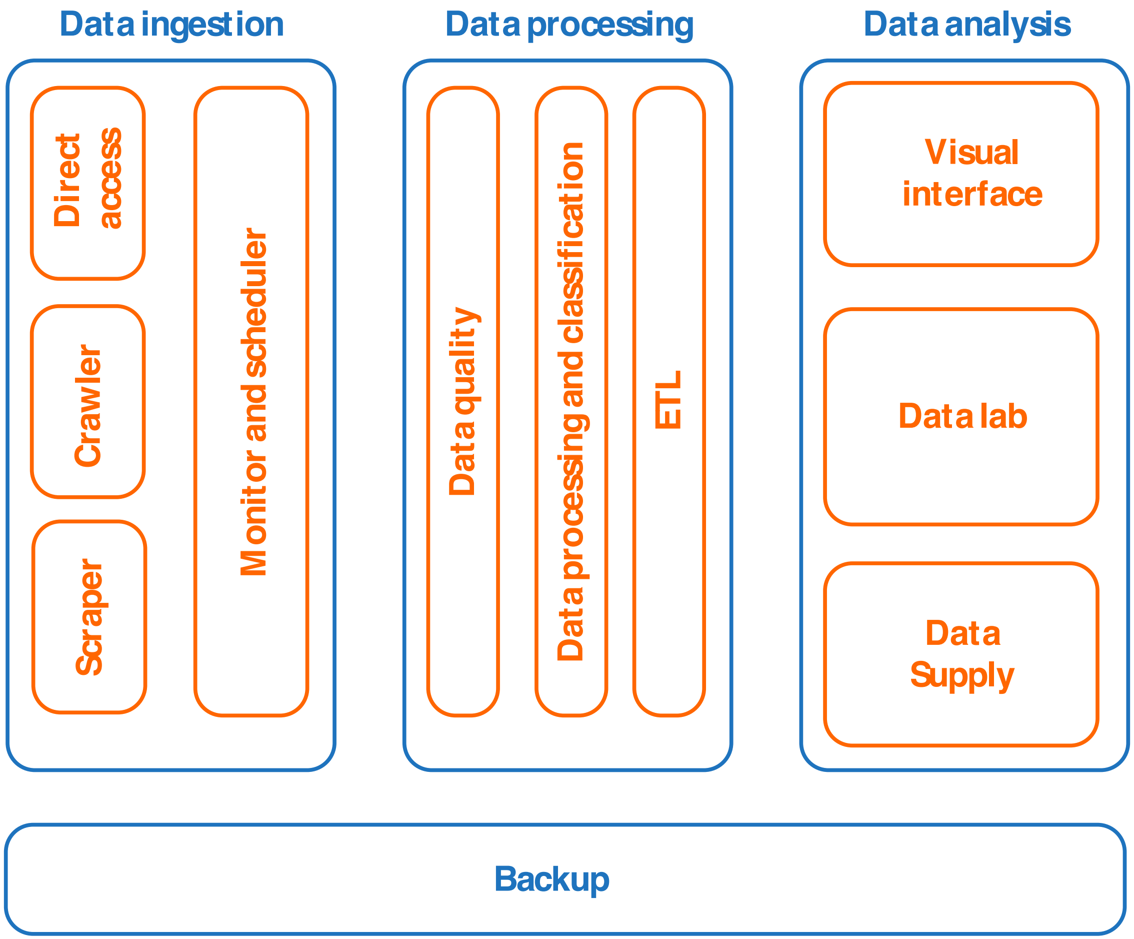
Topics

1. Recap
2. The Data System
    1. **The functional architecture**
    2. Data ingestion techniques
    3. Data processing pipeline
    4. Classification techniques

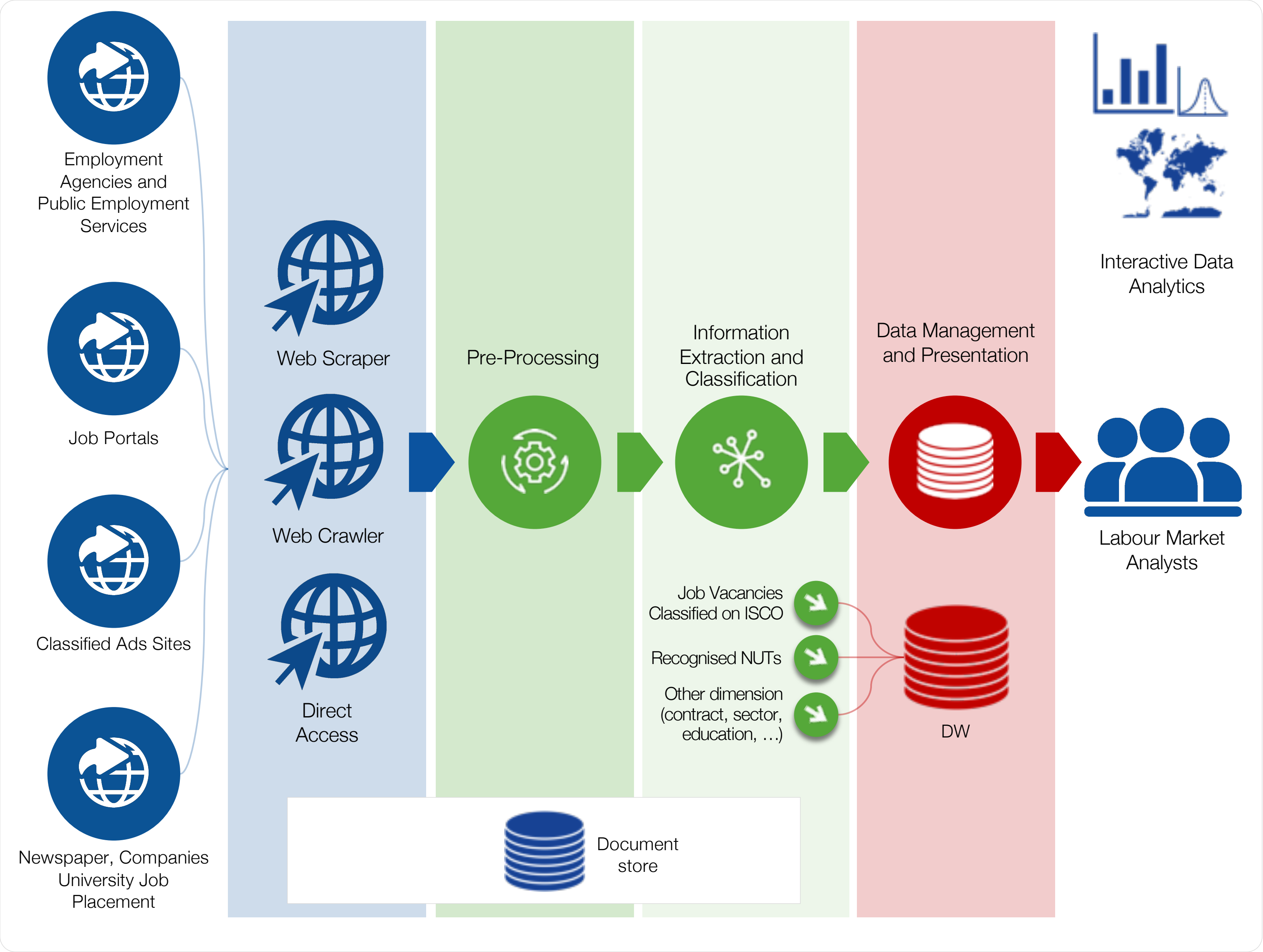
Overall Data Flow



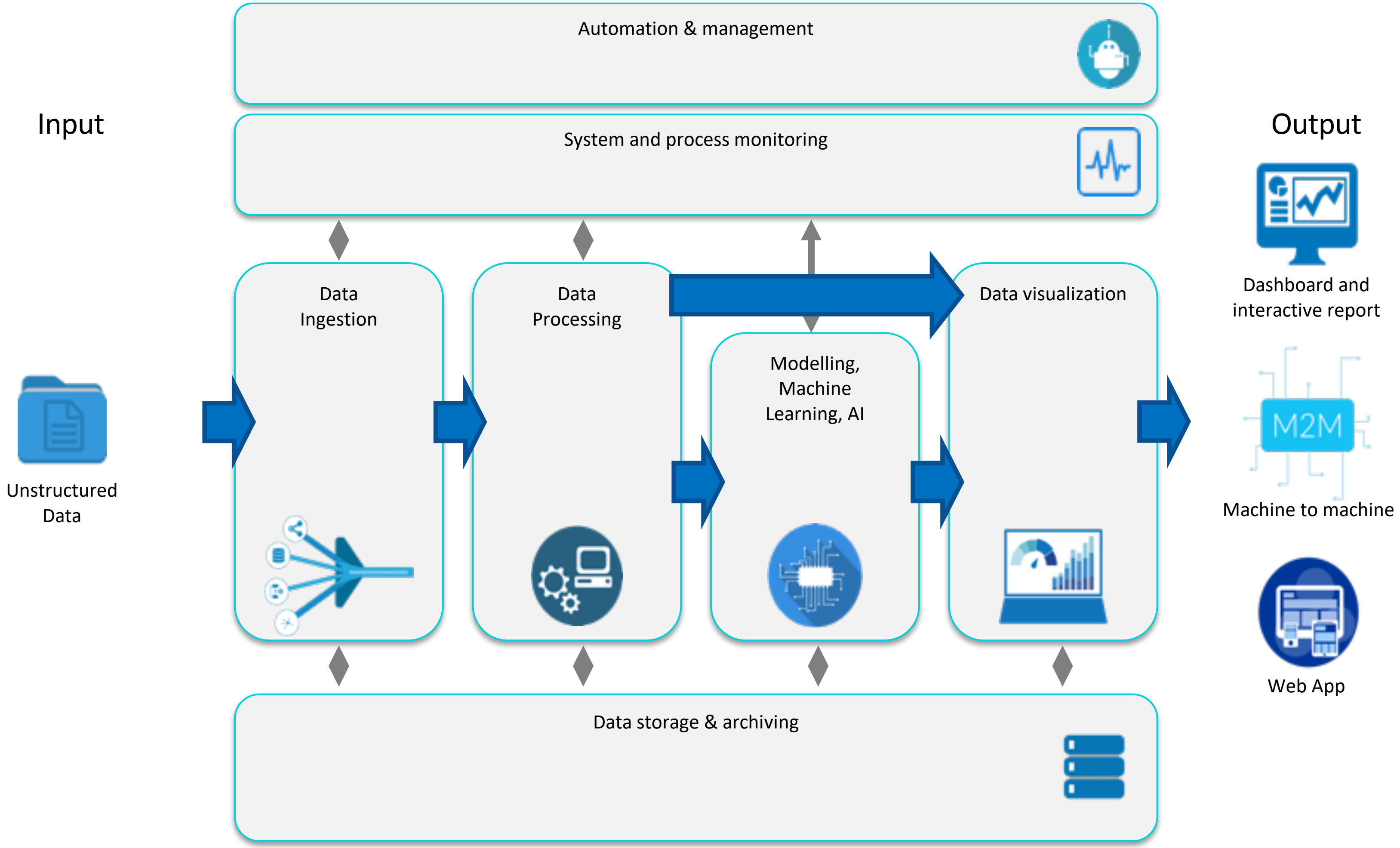
Conceptual architecture



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
Job Vacancies Classified on ISCO
Recognised NUTs
Other dimension (contract, sector, education, …)
DW
Document store
Interactive Data Analytics
Labour Market Analysts

Physical view



Automation & management

System and process monitoring

Input

Output

Unstructured Data

Data Ingestion

Data Processing

Modelling, Machine Learning, AI

Data visualization

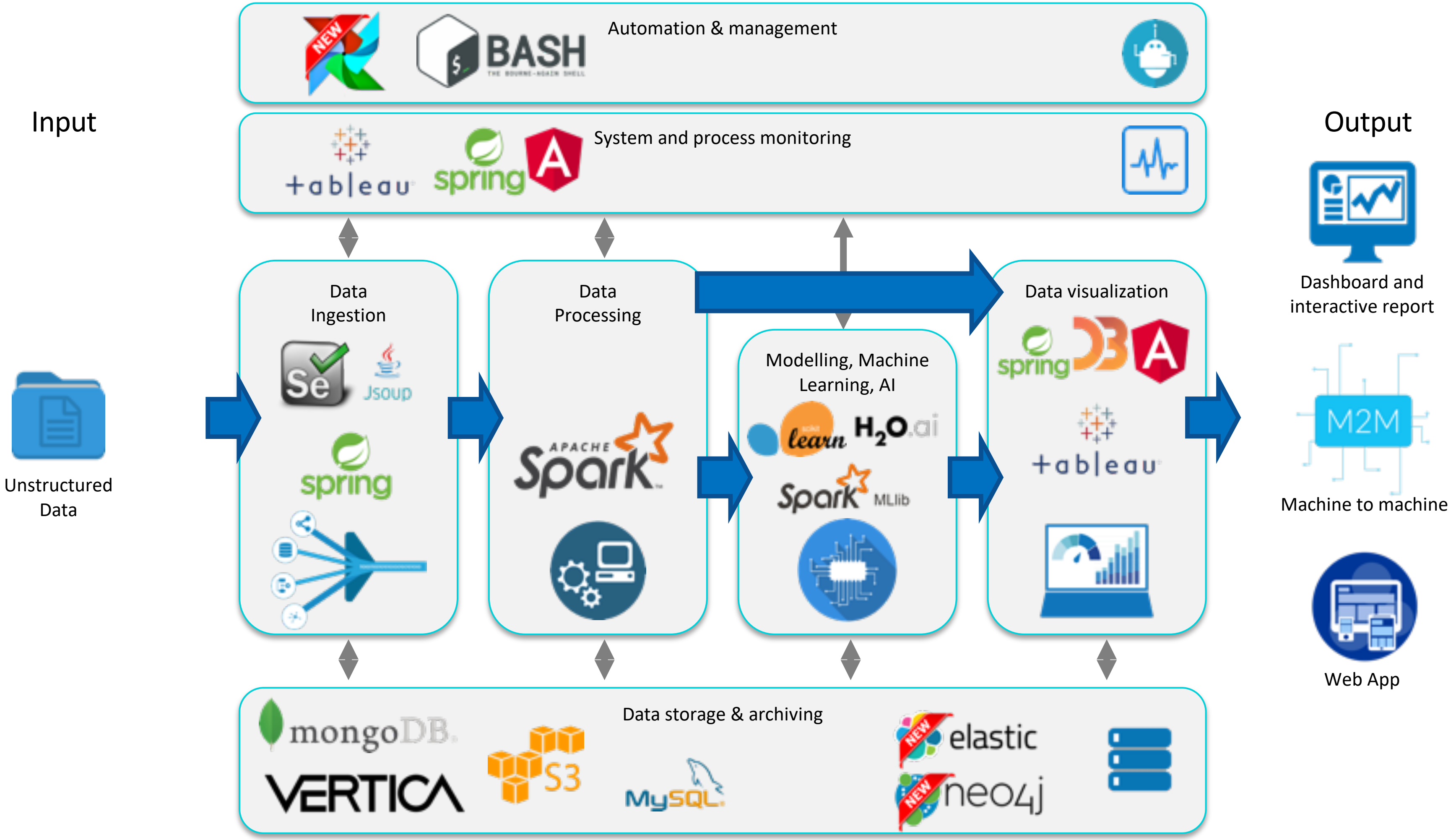
Dashboard and interactive report

Machine to machine

Web App

Data storage & archiving

Technology view


Input

Unstructured Data

Automation & management

System and process monitoring

Data Ingestion

Data Processing

Modelling, Machine Learning, AI

Data visualization

Data storage & archiving

Output

Dashboard and interactive report

Machine to machine

Web App

Topics

1. Recap
2. The Data System
    1. The functional architecture
    2. **Data ingestion techniques**
    3. Data processing pipeline
    4. 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:

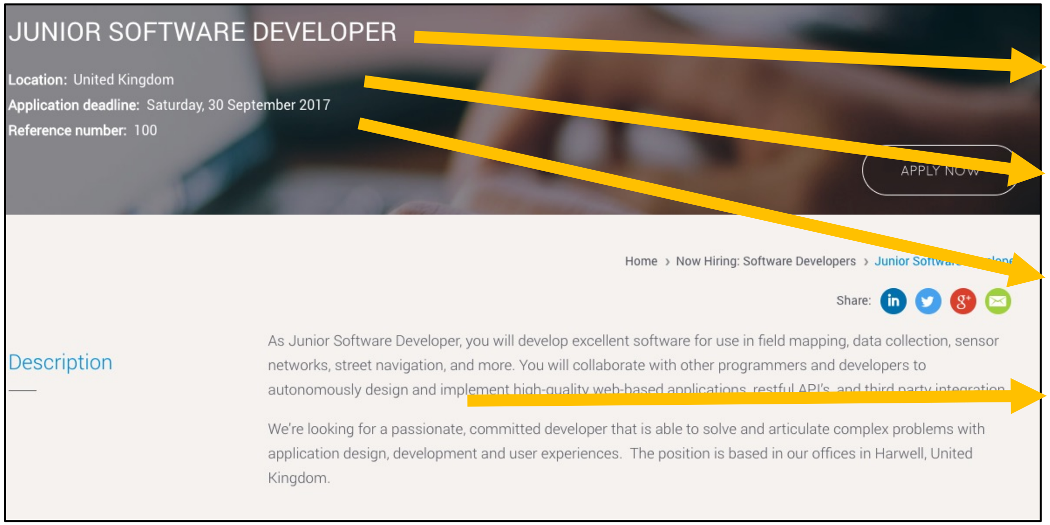
- API
- Scraping
- Crawling

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

Scraping – An example

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



JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom
Application deadline: Saturday, 30 September 2017
Reference number: 100

Home › Now Hiring: Software Developers › Junior Software Developer

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.

Title: Junior Software Developer

Area: United Kingdom

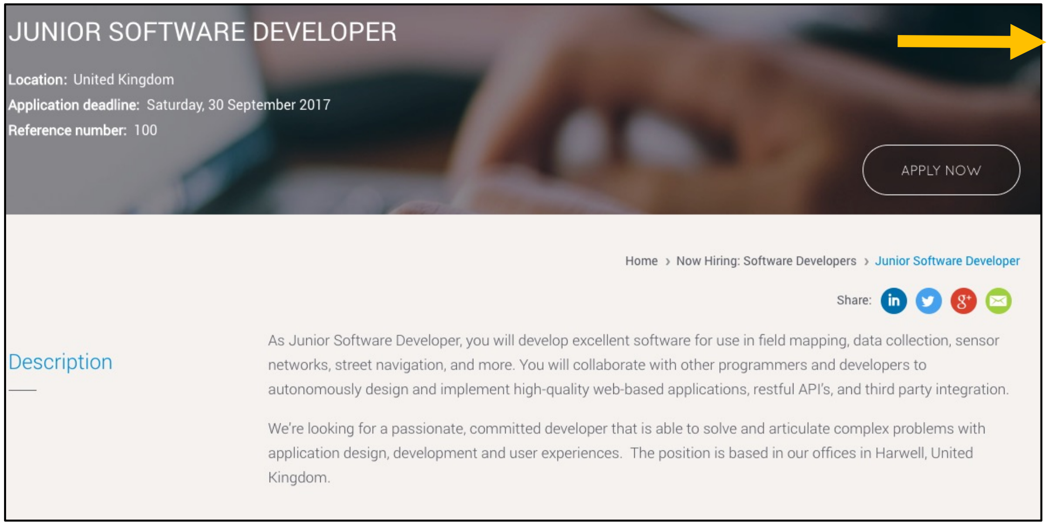
Time: Saturday, 30 September 2017

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)



JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom

Application deadline: Saturday, 30 September 2017

Reference number: 100

APPLY NOW
Home › Now Hiring: Software Developers › Junior Software Developer
Share:

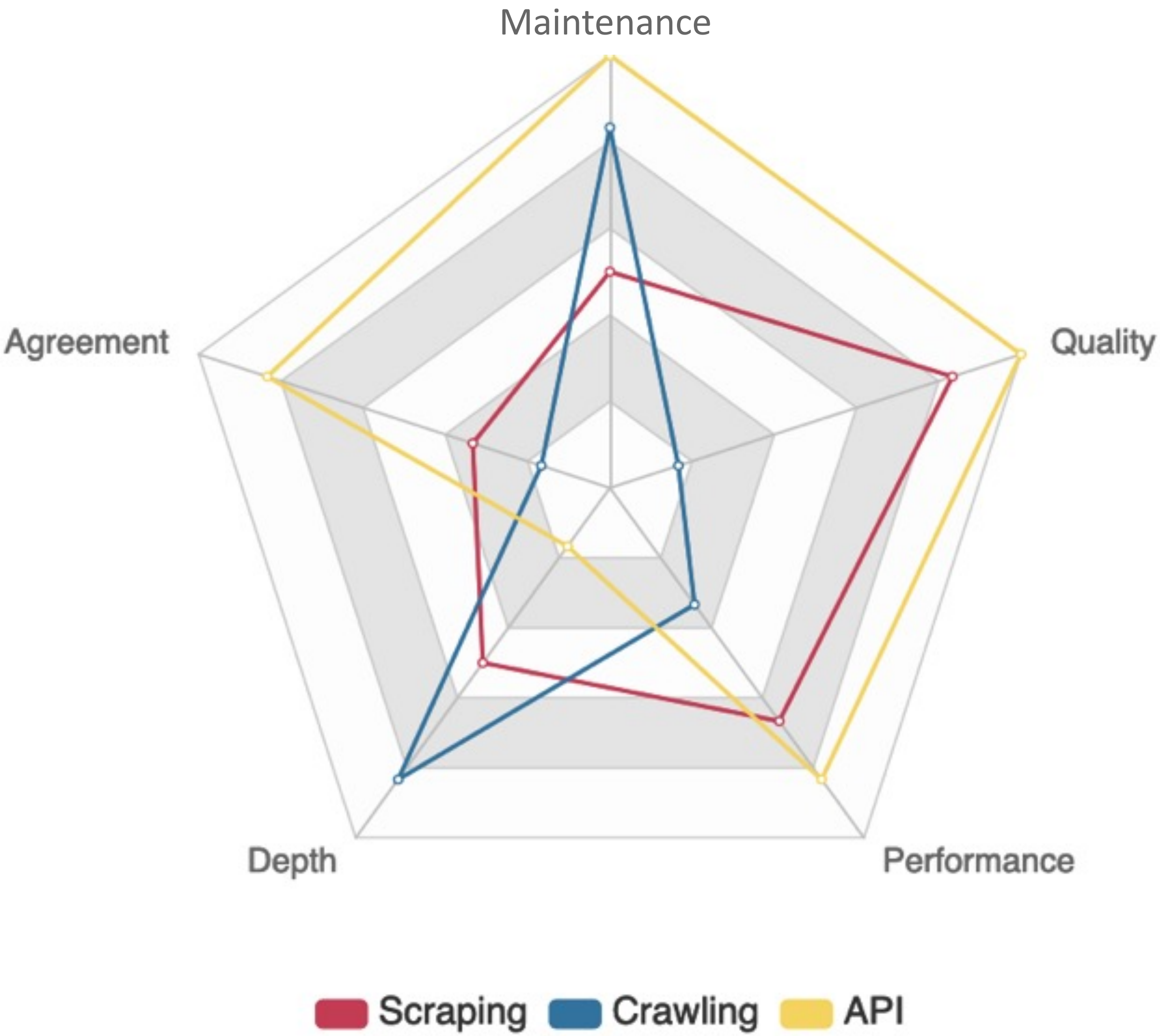
Description

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.

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…
```


Maintenance
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

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

Recap & Keywords

- Landscaping, source selections and augmentation
- Tailored approach
    - API, Scraping, Crawling components
- Focus on quantity
    - Scaling and real-time collecting
- Real-time monitoring of the collected data

Topics

1. Recap
2. The Data System
    1. The functional architecture
    2. Data ingestion techniques
    3. **Data processing pipeline**
    4. 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
    - …

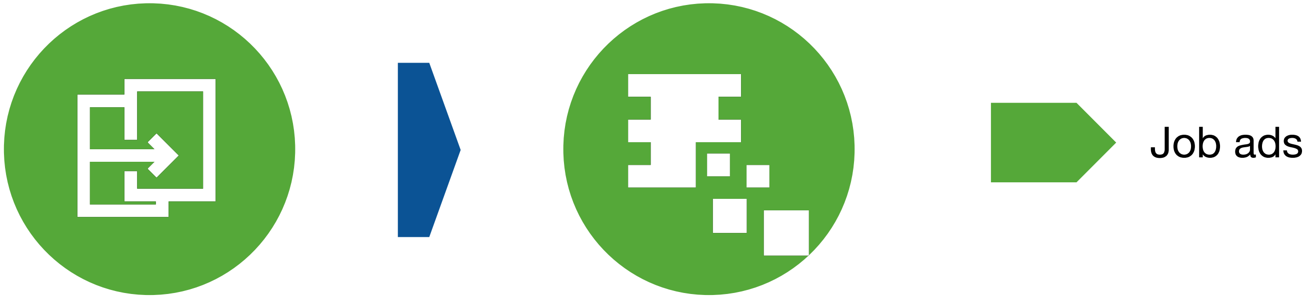
- …Then, we have to detect and handle duplicated OJVs:
  - Generally, a vacancy is posted on multiple portals
  - If we deal with them as distinct, we would overestimate Labour Demand
  - So, we've to detect duplicated OJVs and merge information coming from them in a single one

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

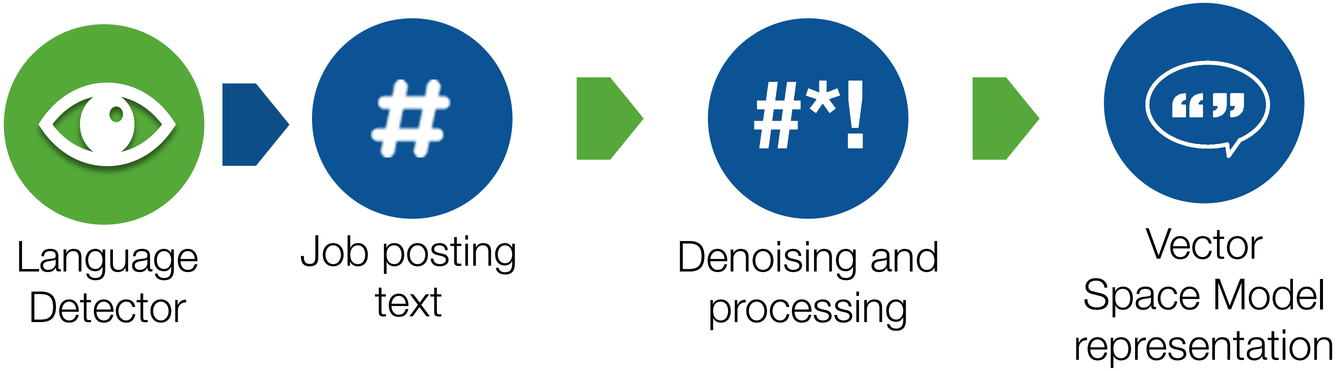
Made on the description (or content) part of the job vacancy.

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.

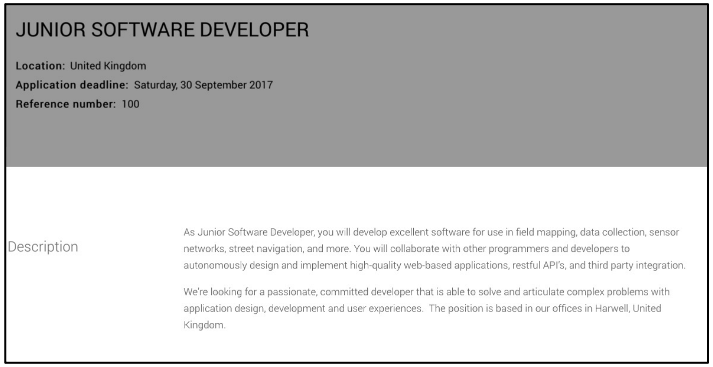


Language Detector

Job posting text

Denoising and processing

Vector Space Model representation



JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom
Application deadline: Saturday, 30 September 2017
Reference number: 100

Description

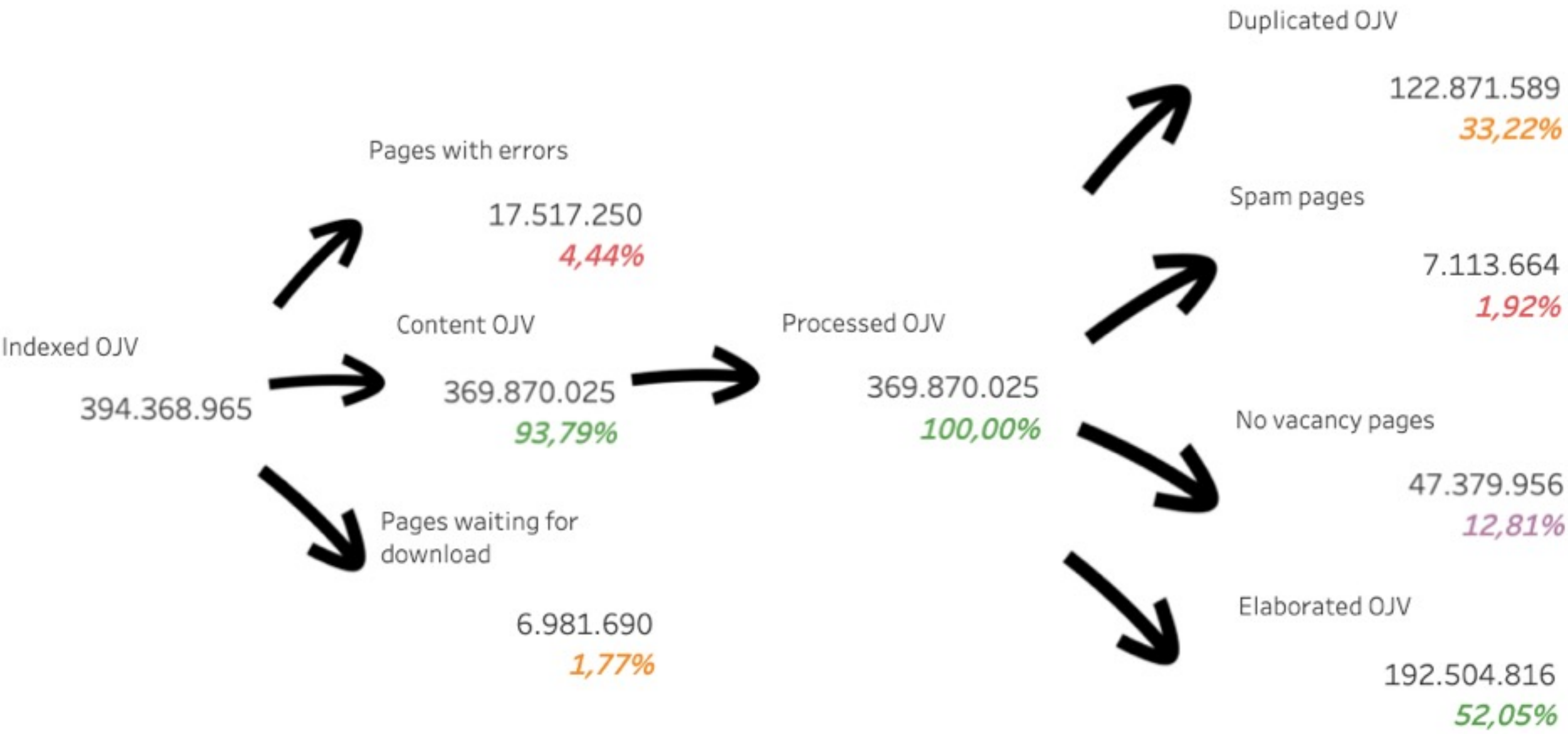
As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful API's,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



Indexed OJV

394.368.965

Pages with errors

17.517.250

4,44%

Content OJV

369.870.025

93,79%

Pages waiting for download

6.981.690

1,77%

Processed OJV

369.870.025

100,00%

Duplicated OJV

122.871.589

33,22%

Spam pages

7.113.664

1,92%

No vacancy pages

47.379.956

12,81%

Elaborated OJV

192.504.816

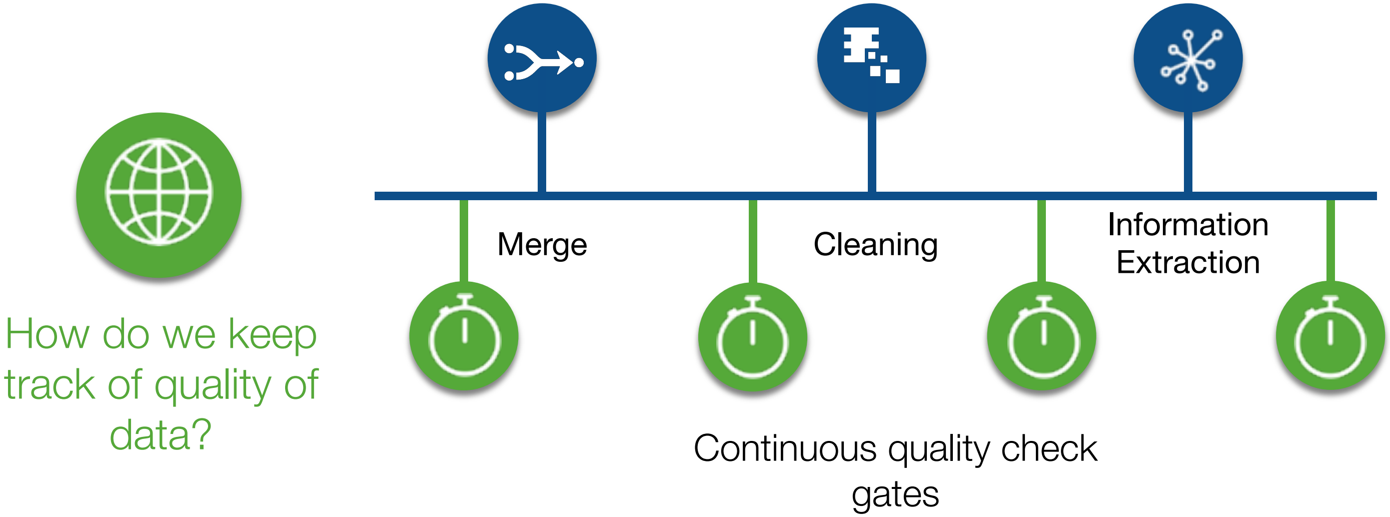
52,05%

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


Merge
Cleaning
Information Extraction
How do we keep track of quality of data?
Continuous quality check gates

Recap & Keywords

- Focus on quality
    - How remove noise?
    - Deduplication activities
- Languages challenge
    - Tailored component for each language
- Track of quality of data
    - Continous quality check and gates

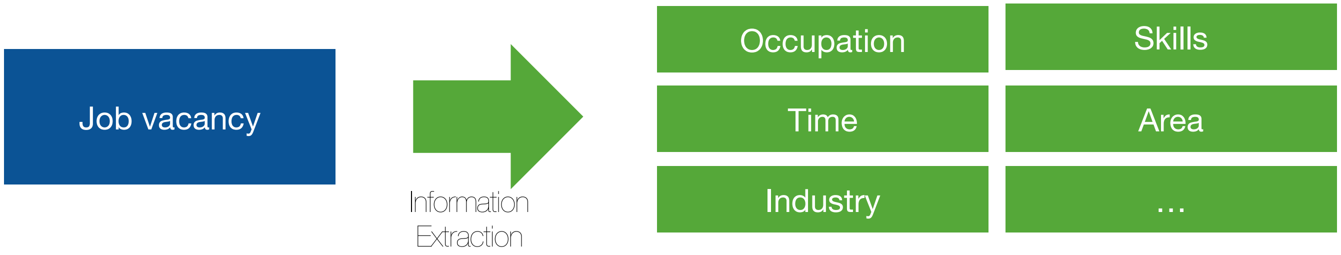
Topics

1. Recap
2. The Data System
    1. The functional architecture
    2. Data ingestion techniques
    3. Data processing pipeline
    4. **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



Job vacancy

Information Extraction

Occupation

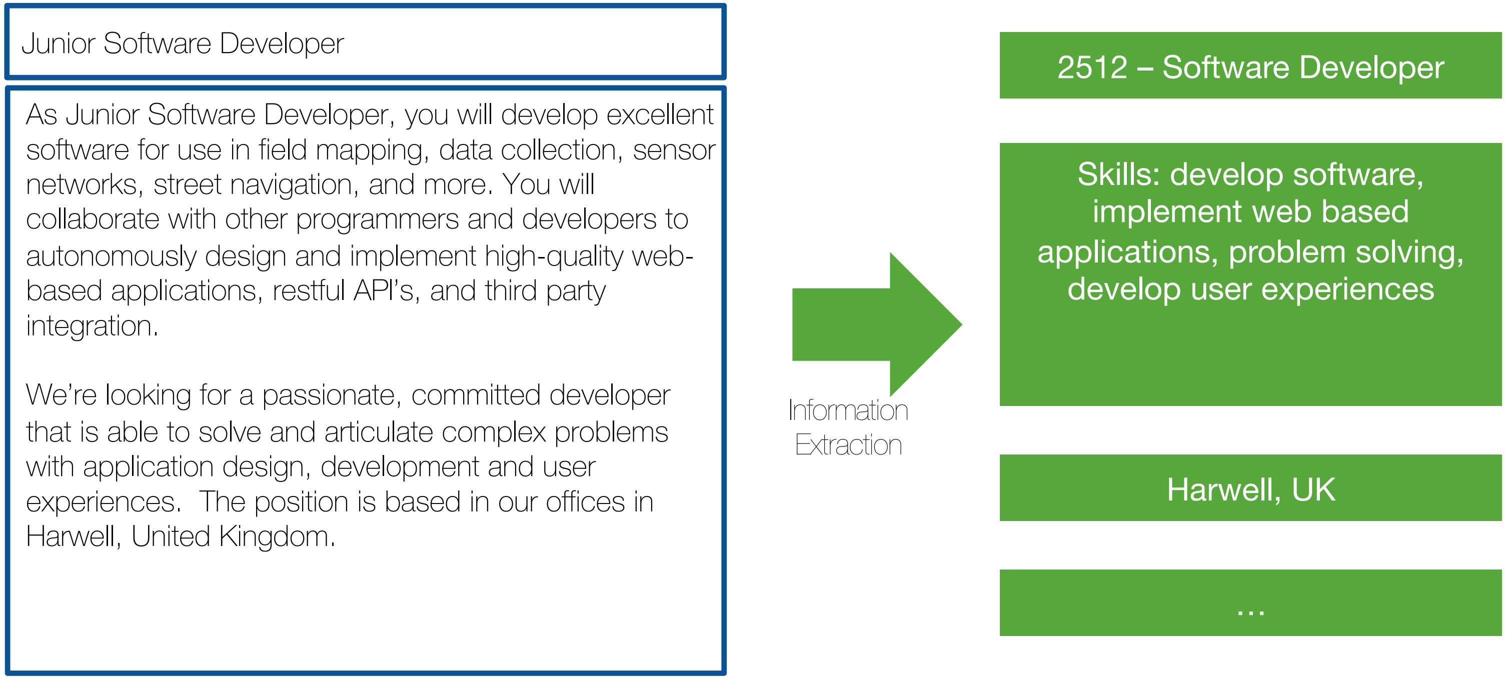
Skills

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

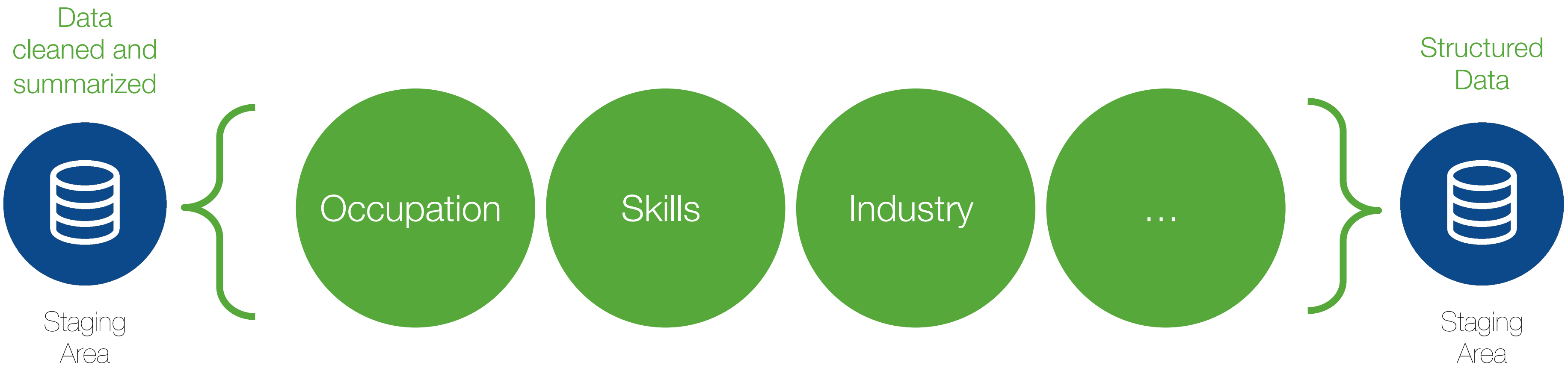
Harwell, UK

…

Information Extraction and Classification
Real Time Labour Market Intelligence

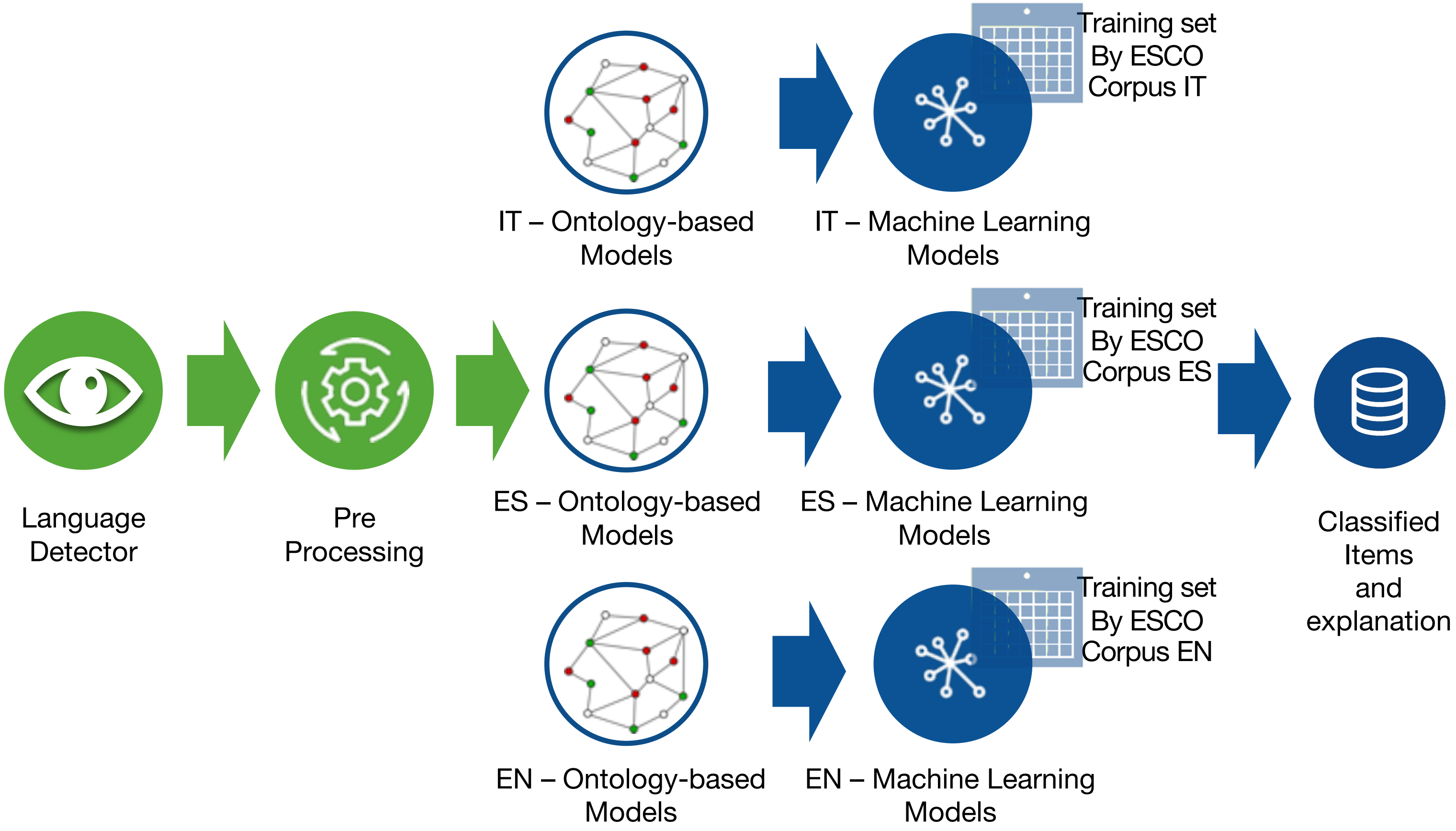
Information Extraction is an area of natural language processing that deals with finding factual information in free text.

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



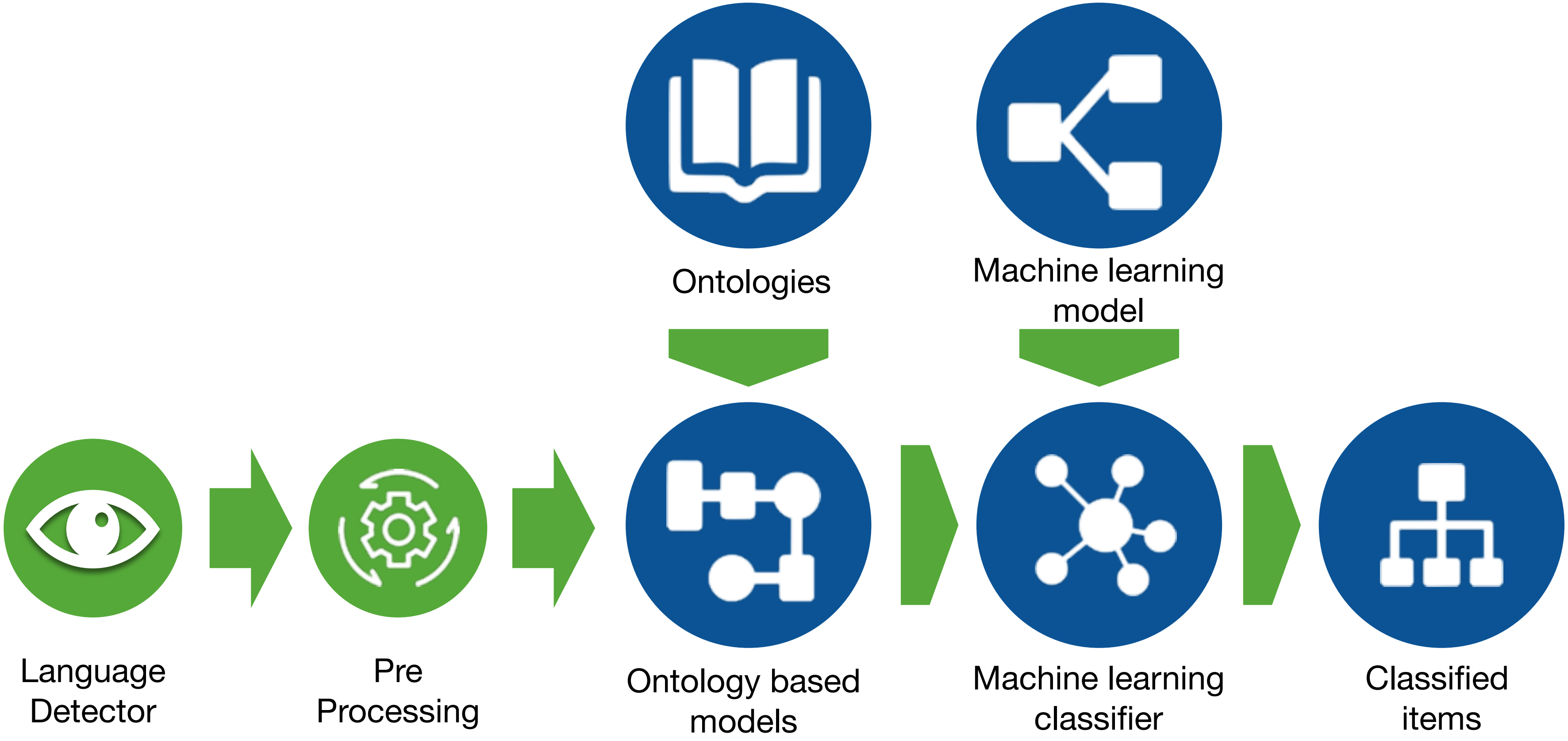
Machine Learning → Ontology based learning, supervised learning and unsupervised learning, etc.

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 based

String similarity measures operate on string sequences and character composition.

Jaro-Winkler, Jaccard, Cosine similarity

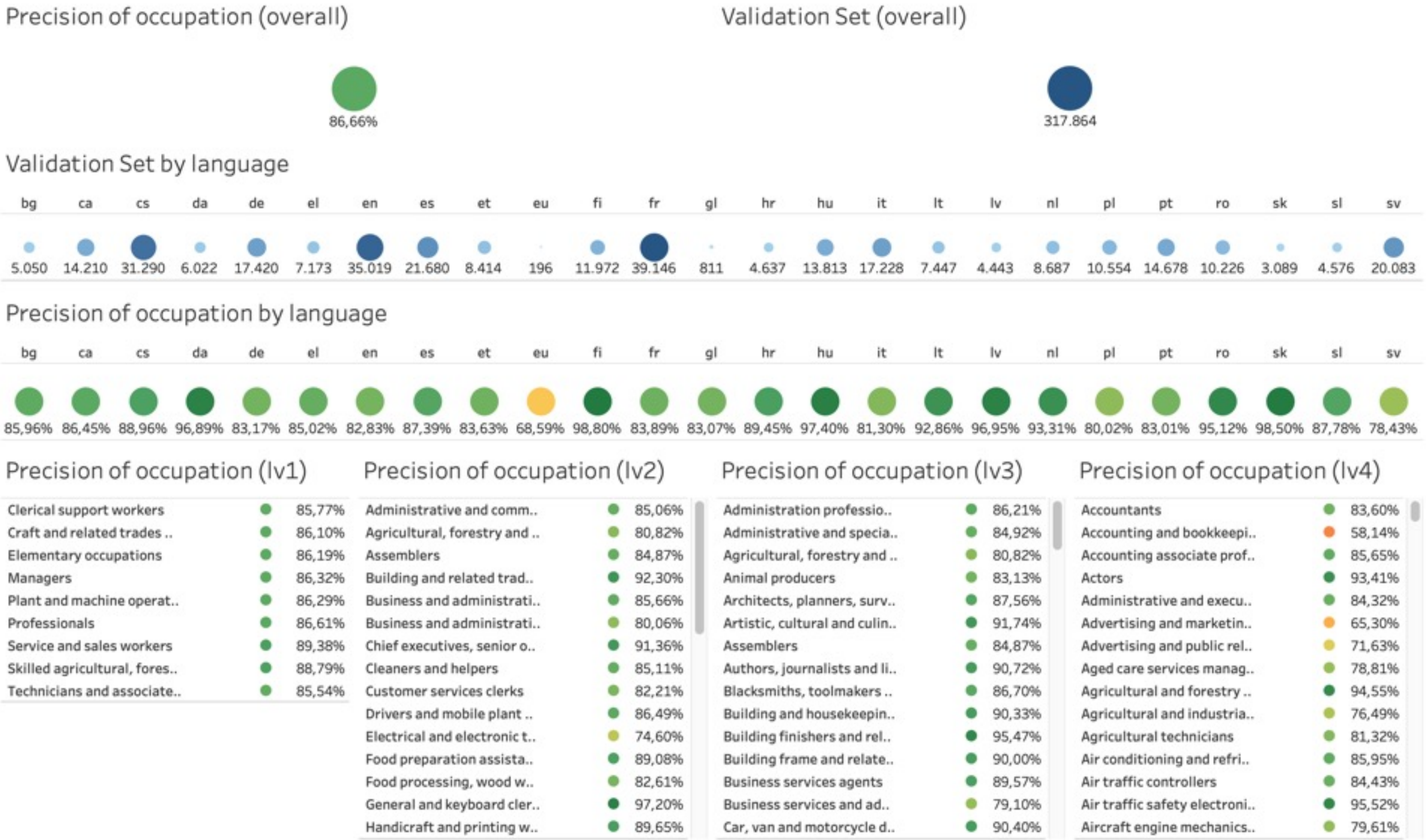
Corpus based

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%

Recap & Keywords

- Focus on summarization
  - How summarize data and improve our data analysts results?
- Link to standard taxonomies
  - Compare OJVs data with other sources
- Gold-set challenges (cardinality, quality and diversity)
- Mixed approaches
  - Machine learning
  - Ontology based learning
  - Text similarity and Information extraction techniques
- Model Life-Cycle