Big Data for Labour Market Information

Session 5
Machine learning, AI algorithms

Alessandro Vaccarino – Fabio Mercorio

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

Milan, 21-22 November 2019

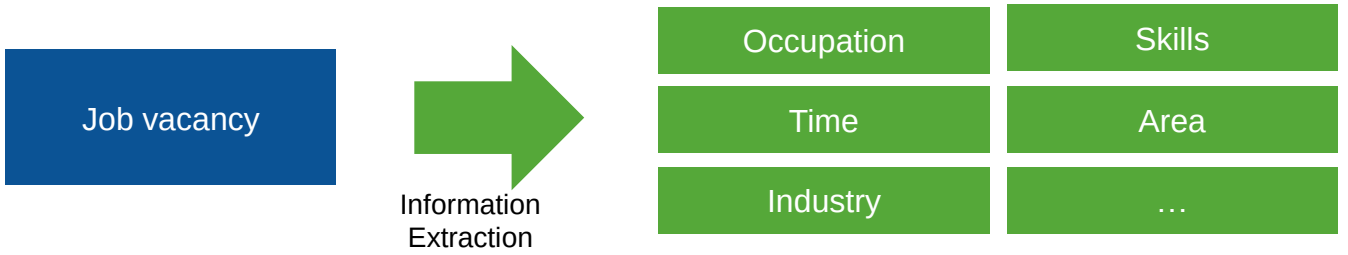
Topics

1. **Data Classification**
2. Model Life-Cycle

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.

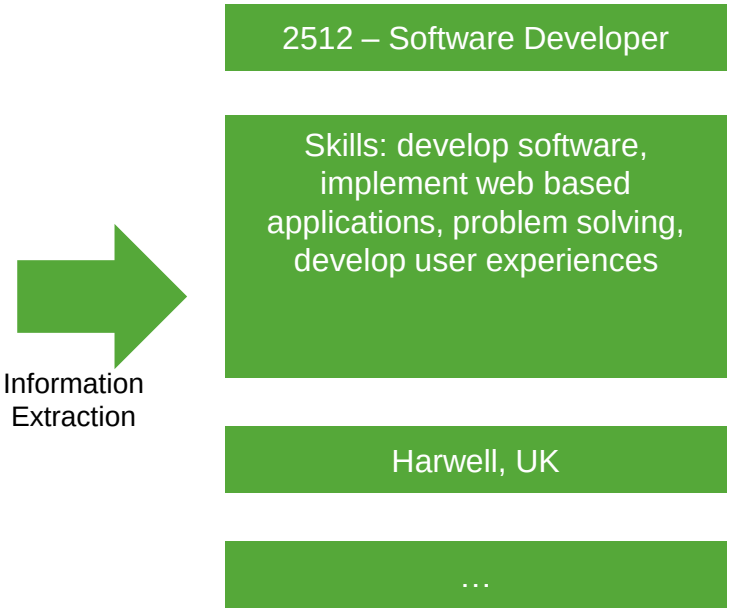
Data Classification - An example



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.

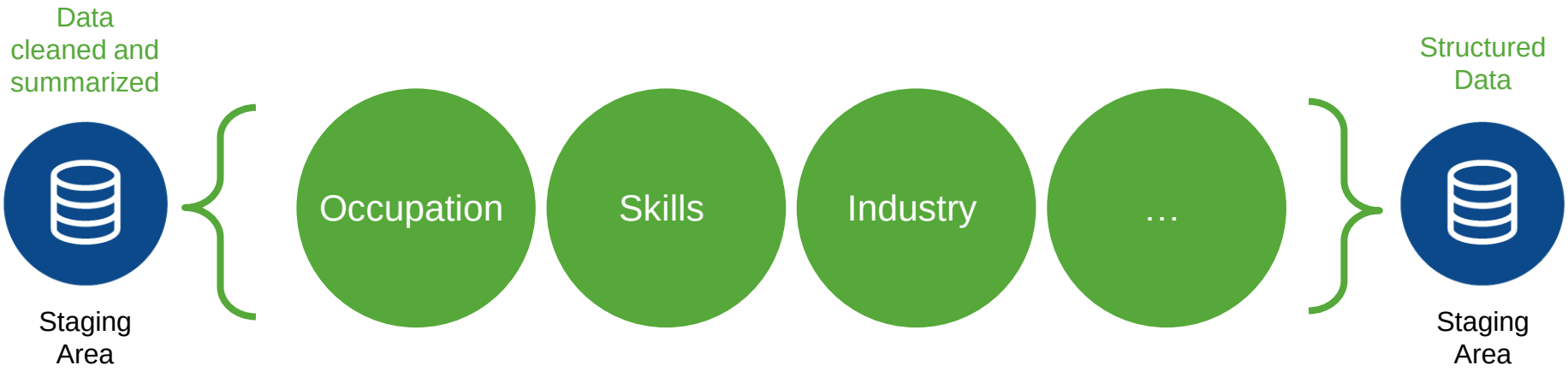
We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.



Information Extraction and Classification
Real Time Labour Market Intelligence

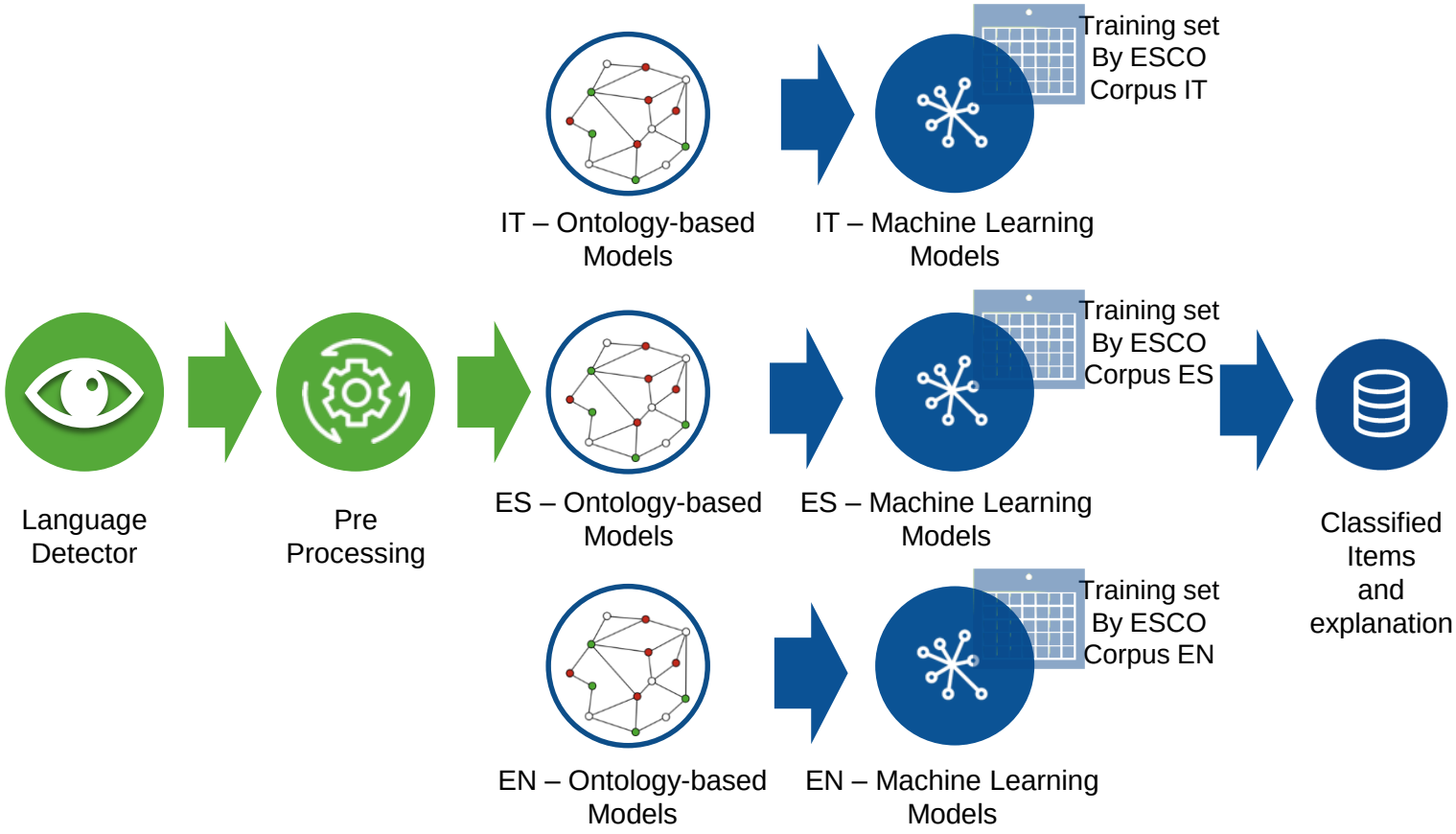
Information Extraction is an area of natural language processing that deals with finding factual information in free text.

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



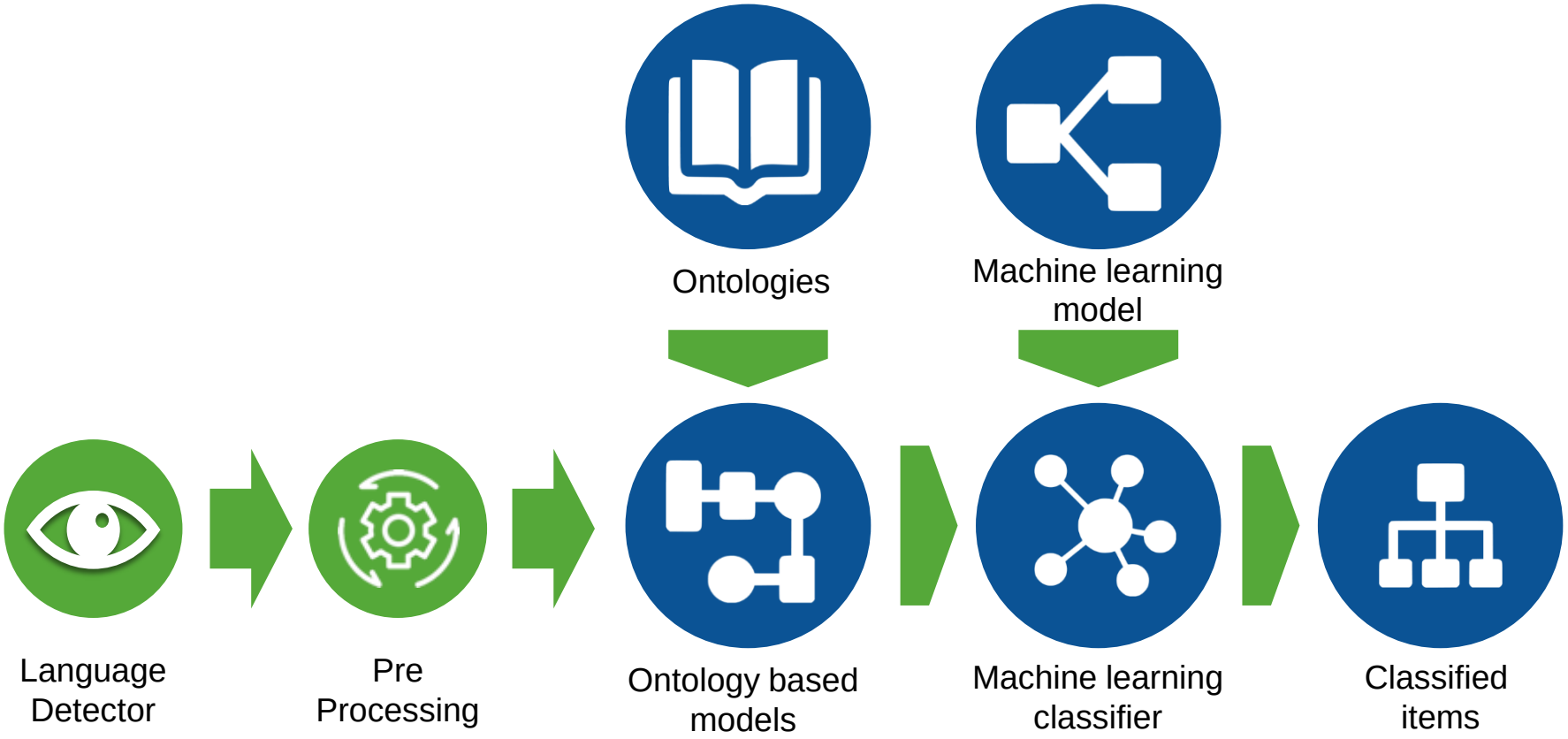
Machine Learning → Ontology based learning, supervised learning and unsupervised learning, etc.

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 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

Sector

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

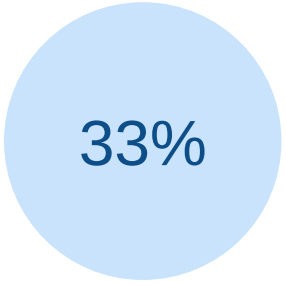
NHS England leads the National Health Service (NHS) in England. We set the priorities and direction of the NHS and encourage and inform the national debate to improve health and care. We want everyone to have greater control of their health and their wellbeing, and to be supported to live longer, healthier 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

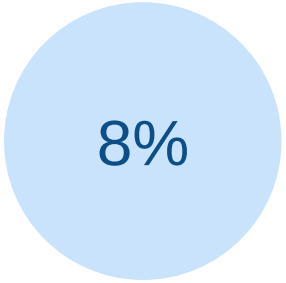
The purpose, to lead the day to day activities to achieve timely stock delivery whilst ensuring that both relevant maintenance standards are achieved and passenger environment activities are enabled. Roles and responsibilities include but are not limited to: Daily delivery of the fleet into service, 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

Custom taxonomy

Structured data

20%

Record linkage with taxonomy

20%

Record not linked (unclassified)

80%

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

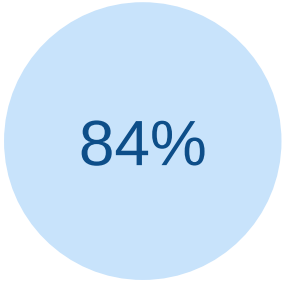
NHS England leads the National Health Service (NHS) in England. We set the priorities and direction of the NHS and encourage and inform the national debate to improve health and care. We want everyone to have greater control of their health and their wellbeing, and to be supported to live longer, healthier 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%

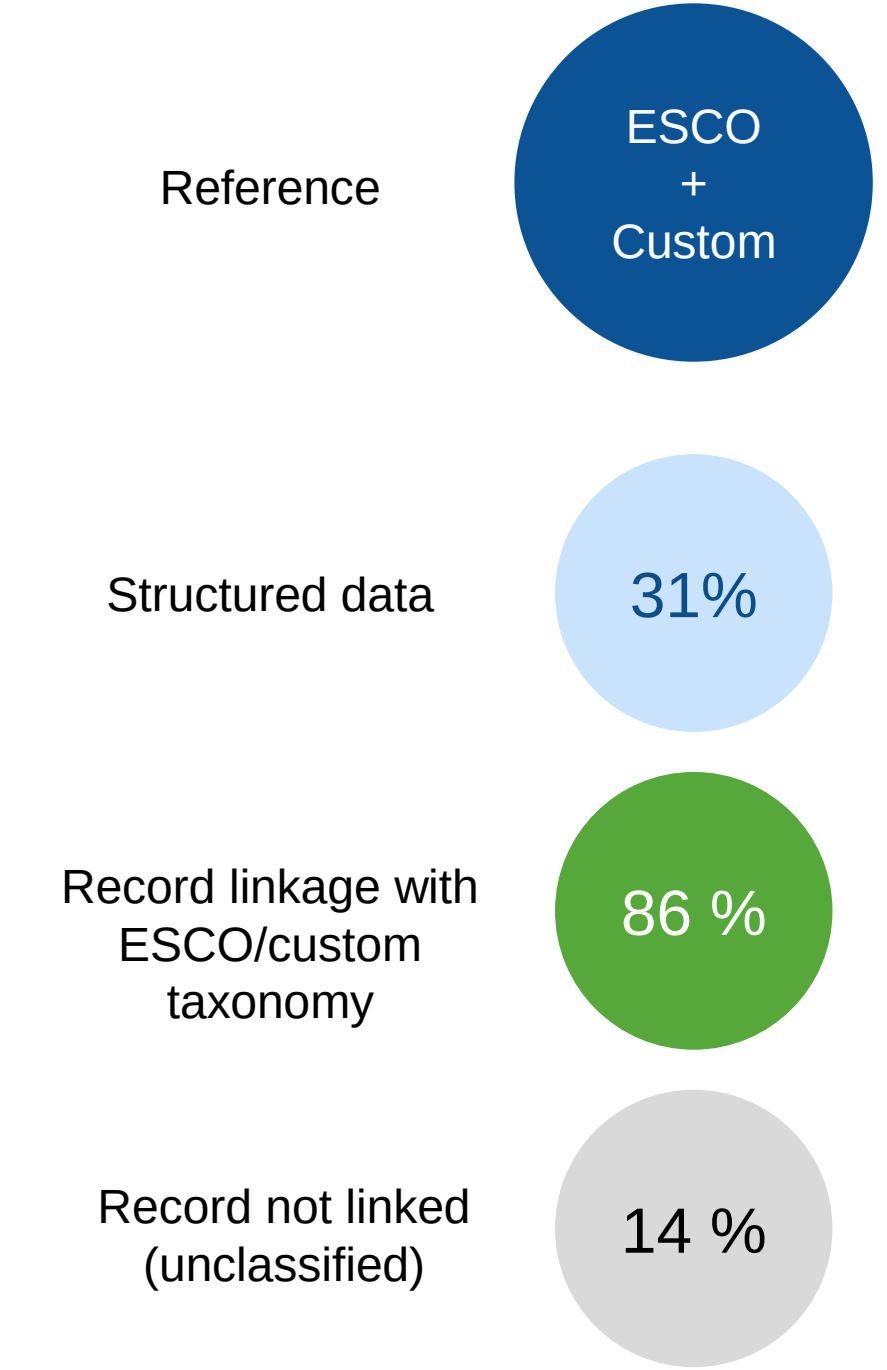
Skill

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**
- ….

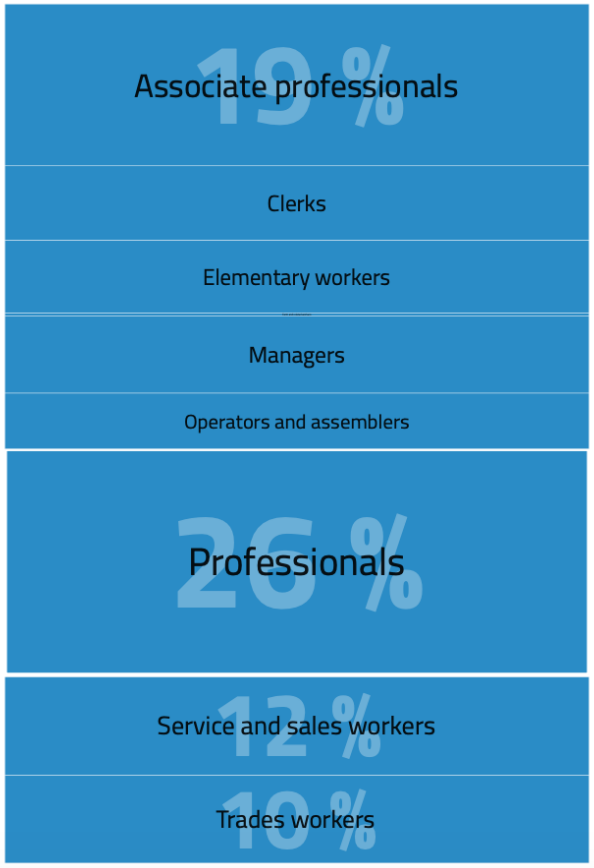


Can you imagine some big data analysis collected so far?

in EU

in Broad occupations

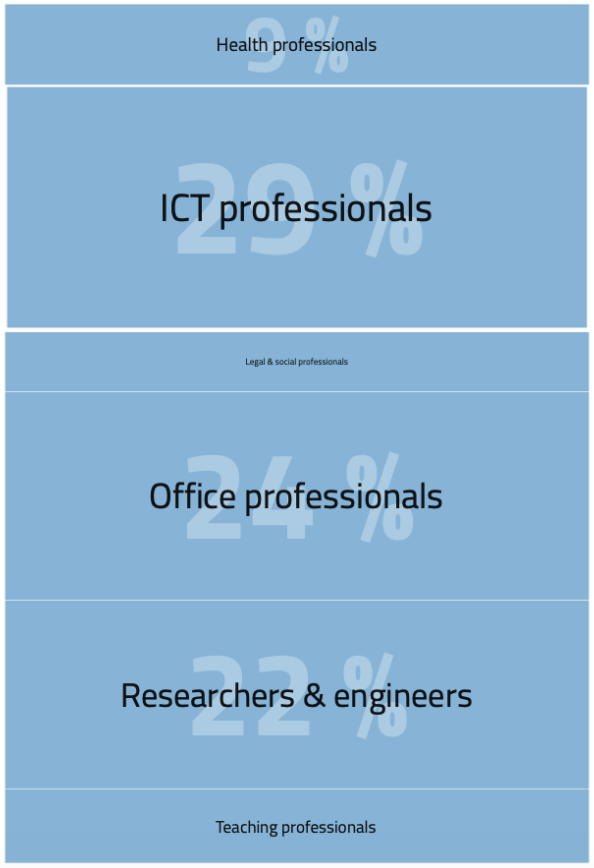
Professionals : 26.3%



in EU

in Professionals

Researchers & engineers : 5.8%



in EU

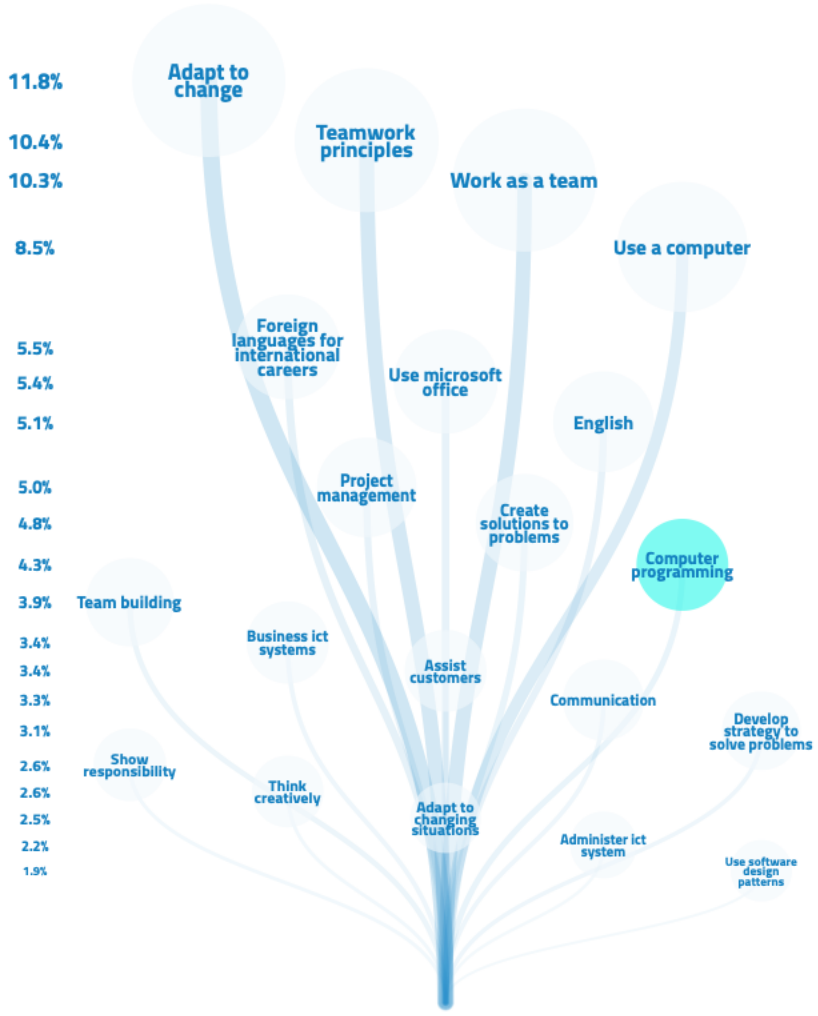
in ICT professionals

Software and applications developers and analysts : 6.5%

SKILLS PANORAMA

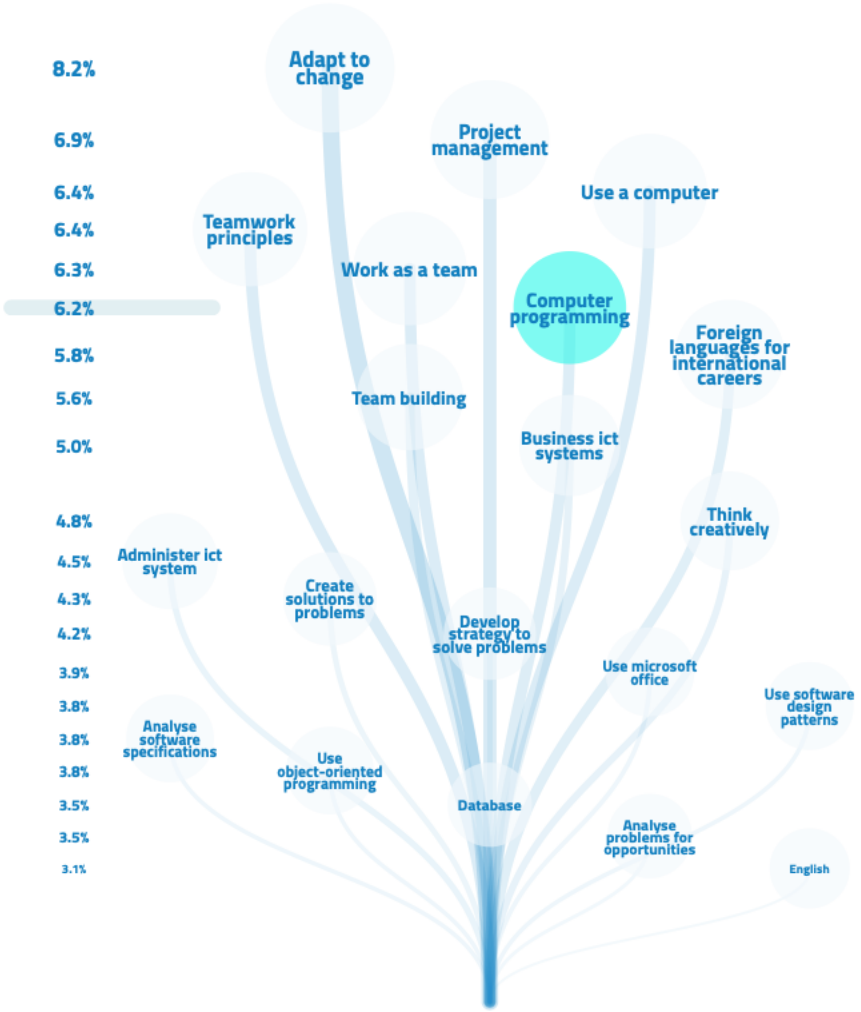
Inspiring choices on skills and jobs in Europe

Most requested Skills in EU in Professionals



Most requested Skills in EU in ICT professionals

Computer programming : 6.2%



SKILLS PANORAMA

Inspiring choices on skills and jobs in Europe

LABOUR MARKET DIGITAL MONITOR

OCCUPATION OF THE MONTH

IN THE LAST 12 MONTH

INFORMATION AND COMMUNICATIONS TECHNOLOGY OPERATIONS TECHNICIANS (ESCO cod. 3511)

Support the day-to-day processing, operation and monitoring of ICT systems, peripherlas, hardware, software and related computer equipment.

OCCUPATION DETAILS

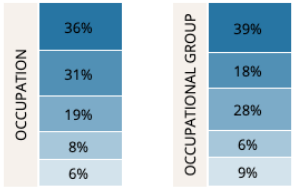
6,405
VACANCIES

37 %
PERMANENT CONTRACTS

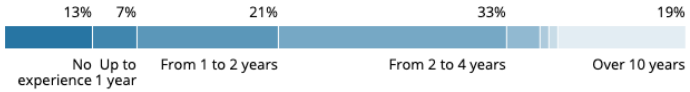
80 %
EXPERIENCE OVER 2 YEARS

CONTRACT TYPOLOGIES

- Temporary
- Permanent
- Self Employment
- Unknown
- Internship

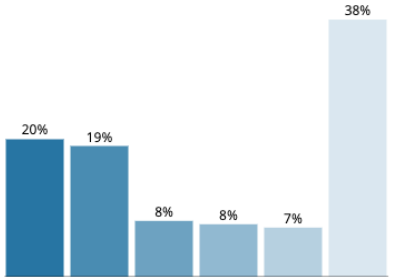


EXPERIENCE

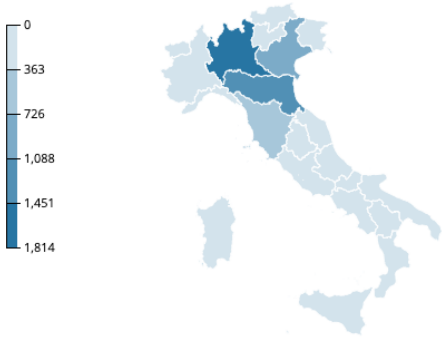


FIRST 5 SUBSECTORS

- Manufacture of electrical equipment
- Manufacturing (General)
- Computer programming, consultancy and related activities
- Information and communication (General)
- Administrative and support service activities (General)
- Others



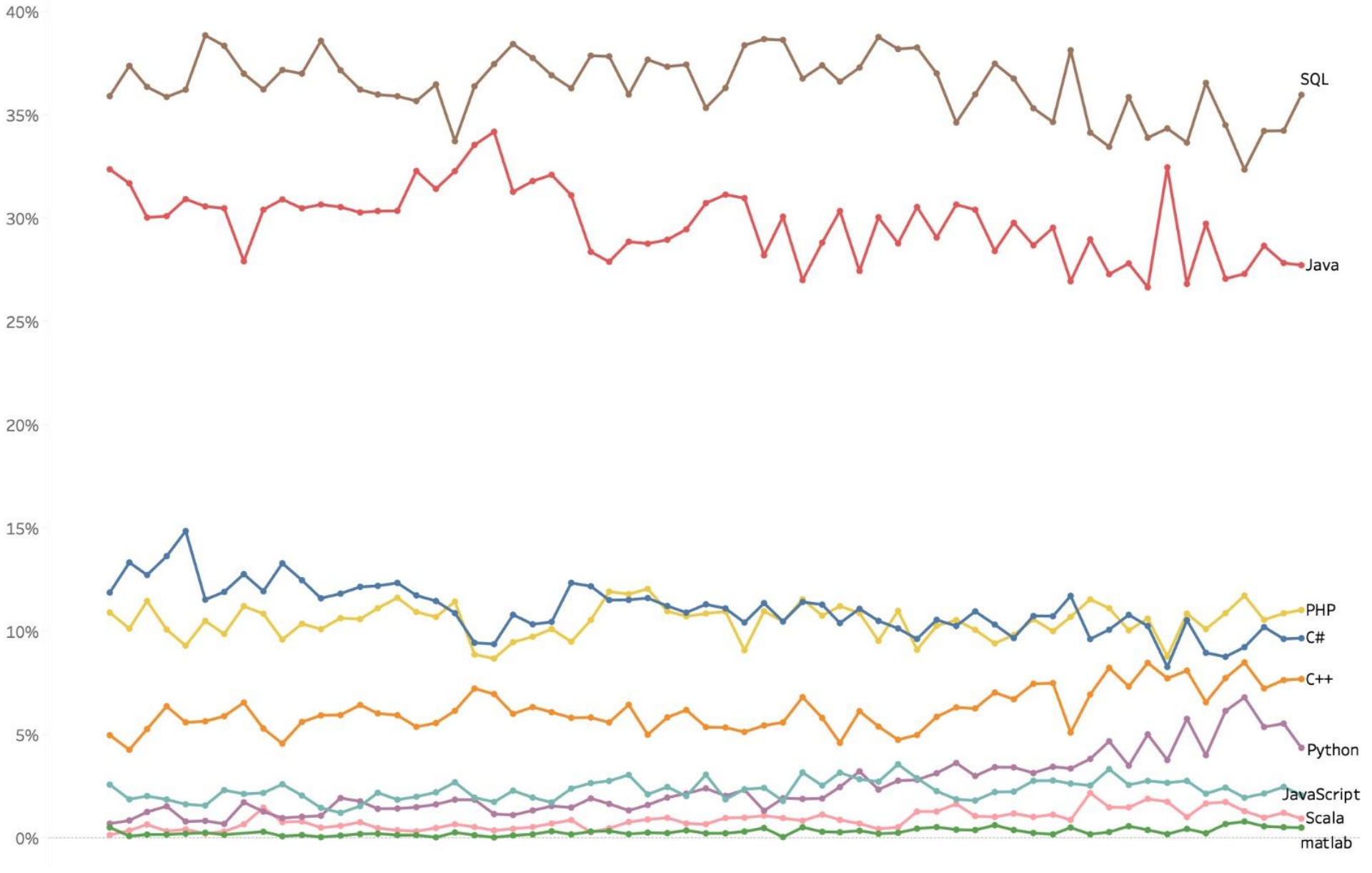
VACANCIES DISTRIBUTION



RELEVANCE

HARD SKILLS	SOFT SKILLS
automation technology	English
CAE software	adapt to change
use software design patterns	analyse problems for opportunities
analyse software specifications	develop strategy to solve problems
use technical drawing software	delegate activities

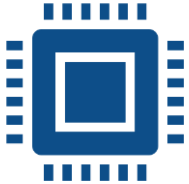
Number of Web Job Vacancy by Programming Language



Data classification
Machine learning recipe

Gold dataset

A training dataset is a dataset of examples used for learning



Classifier

The classifier is the concrete implementation of an algorithm that implements classification

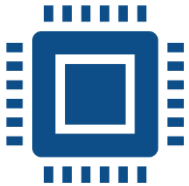
Metric

The metric that you choose to evaluate your machine learning model

Data classification
Occupation

Gold dataset

~150 / 200 k records by each language



Classifier

Ontology based Model + Naive Nayes Classifier

Metric

Weighted Precision

Topics

1. Data Classification
2. **Model Life-Cycle**

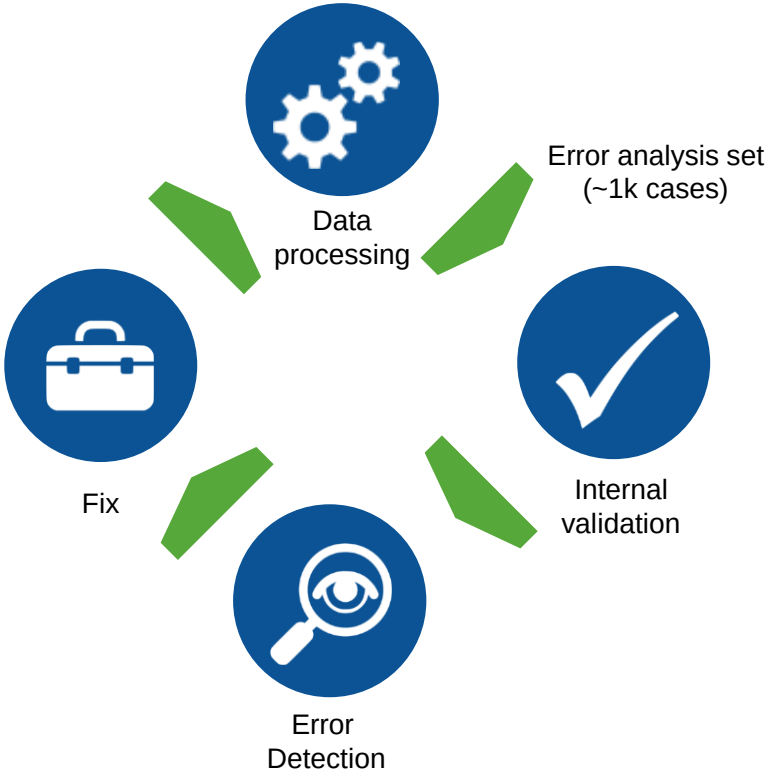
Data classification Challenges: Continous improvement

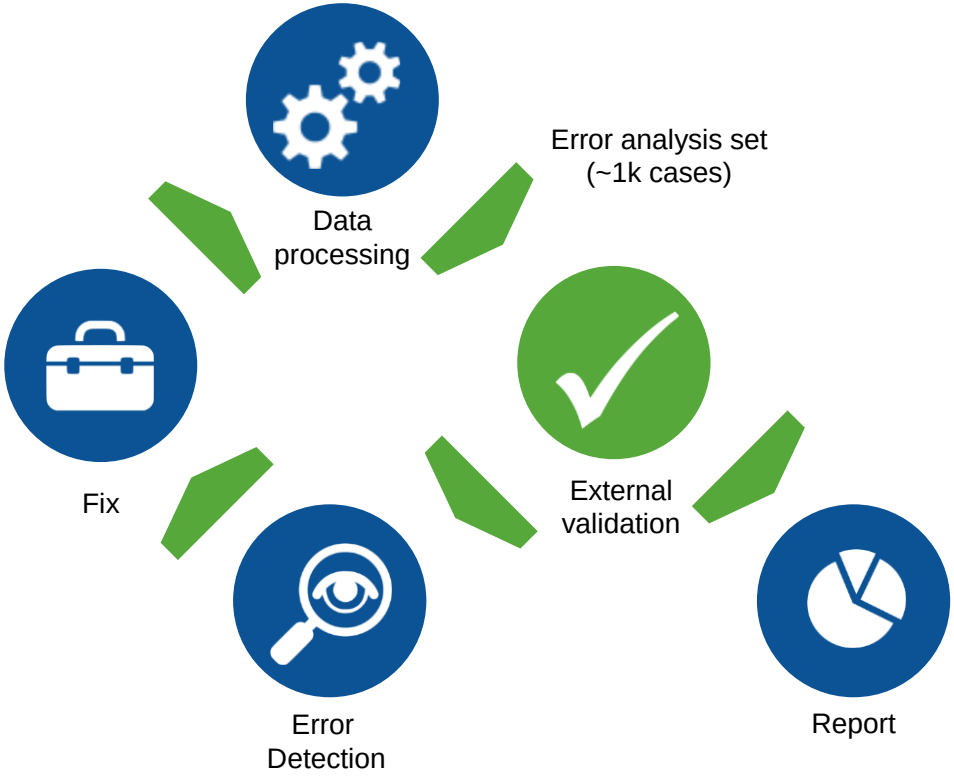
To guarantee a satisfying quality level of classification, we need to keep our models updated

It’s a Model Life-Cycle challenge

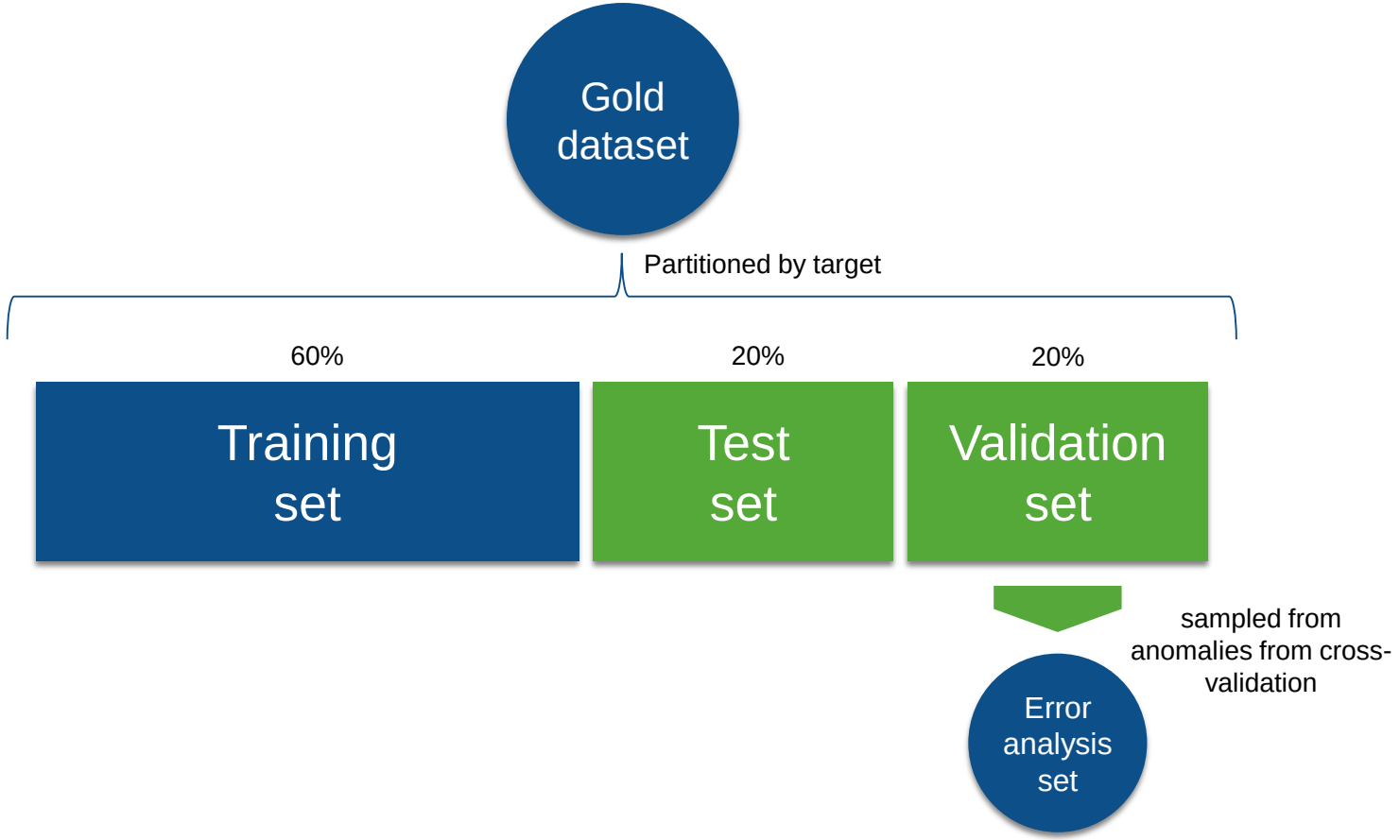
Let’s have a quick deeper look on in…

Model Life-Cycle
Machine Learning diagnostic





Model Life-Cycle
Error analysis set



Model Life-Cycle
Error Analysis

1

Start with a first algorithm, implement and test it

2

Plot learning curves to decide if more data, more features, …

3

Expert validation: manually examine the errors on examples in the cross validation set and try to spot a trend.

4

Apply Expert hints and re-check learning curves

Model Life-Cycle Recap

Get more training examples

Try smaller sets of features

Add new features

Try adding more complex features

Check hyper-parameter of algorithm

More training examples:
Mark with OK/NO the records

Find new labels and new features: fix an association between record and taxonomy

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%