CRISP centro di ricerca interuniversitario per i servizi di pubblica utilità

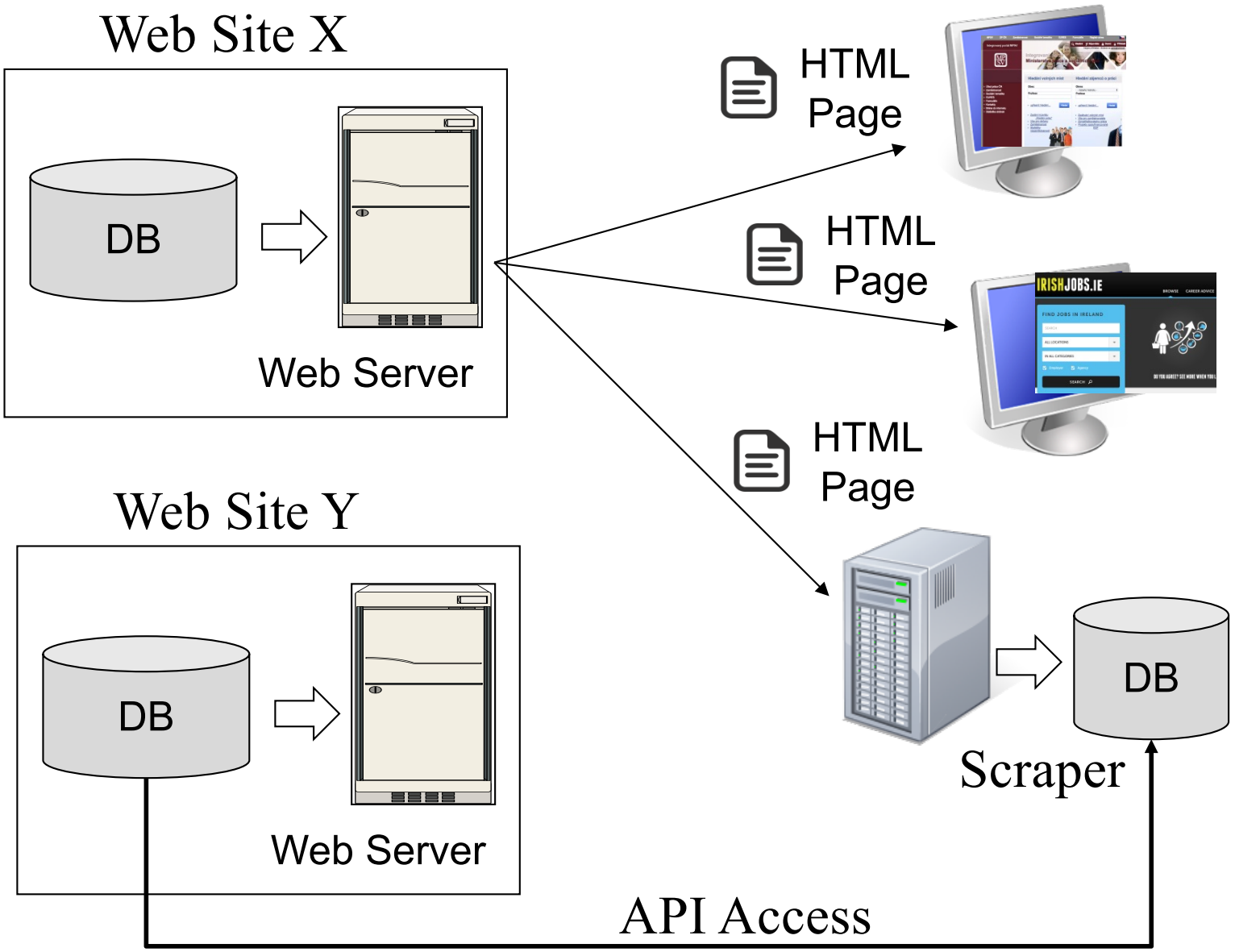
CEDEFOP European Centre for the Development of Vocational Training

Real-time Labour Market Information on Skill Requirements: Feasibility Study and Working Prototype

HOW TO DEAL WITH MILLIONS OF WEB JOB VACANCIES?

- Challenges (focus only on bold items)
  - Identifying web sources [Not shown]
  - **Collecting Job Advertisements from heterogeneous Web Sources**
  - **Extract information**
  - Map site specific data to taxonomies (e.g., geographic locations) [Not shown]
  - **Classify Job Advertisements to the *proper* ISCO (4th digit) code**
  - Extracting skills [Not shown]
  - Analyse and Visualize data for decision making purposes [Not shown]

Web Scraping Scenario

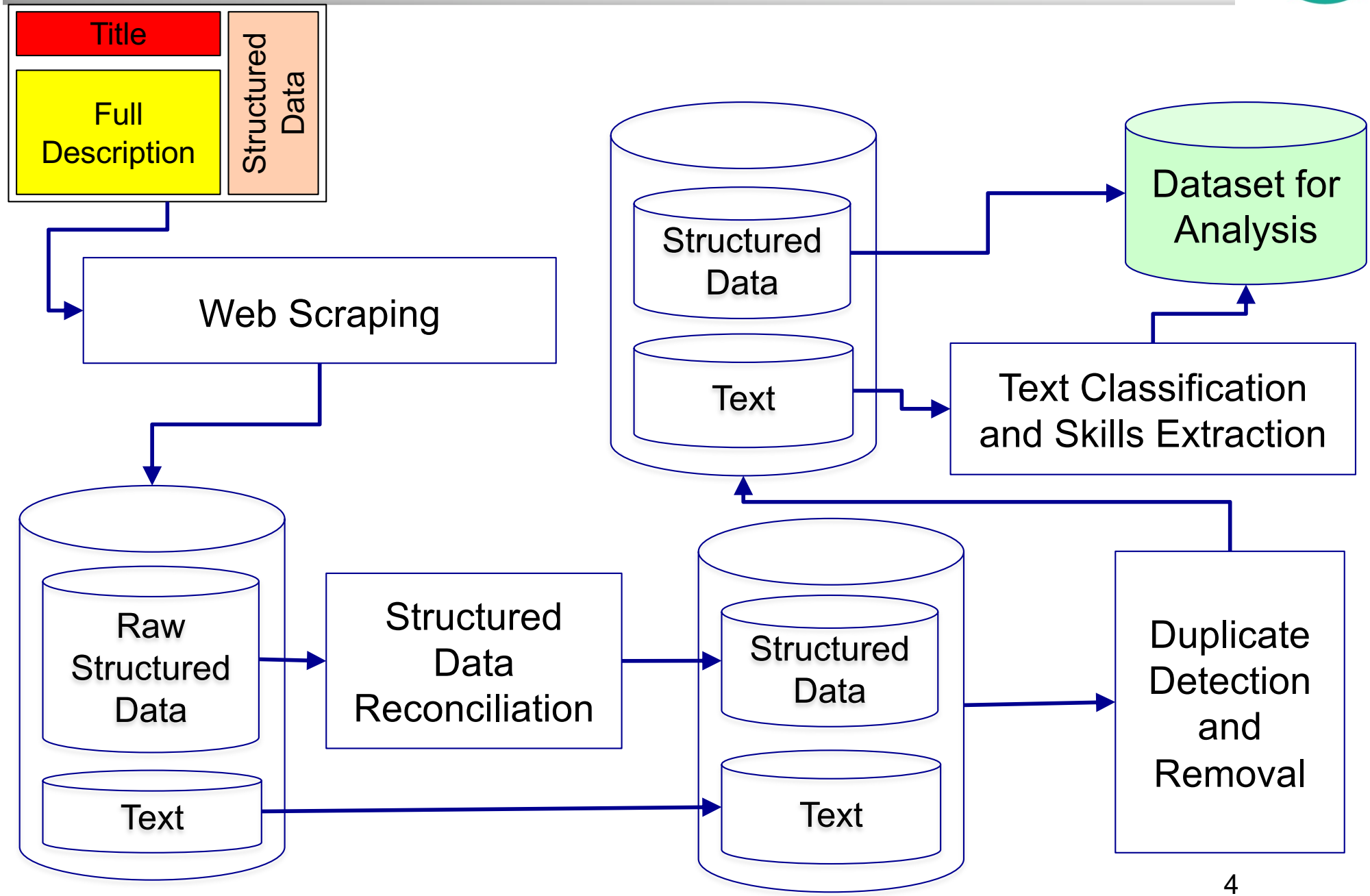


Scraping is resource consuming

To avoid scraping some web sites prefers granting API access to data

API: Application Program Interface

Overall Data Flow



Vacancy Information Extraction

Title

Credit Control Assistant

Location

London

Industries

Accounting and Auditing Services

Job Type

Full Time Permanent

Our client is a well-known group of companies specialising in import and wholesale of consumer of goods. The Credit Control Team needs to expand with the recruitment of an Assistant who can grow with the business. You will be immediately responsible for the following:

- Providing excellent customer service by assisting with the collection of debts for the Credit Control Manager
- Updating and maintaining customer records
- Taking payments
- Dealing with customer queries and issuing copy invoices as required

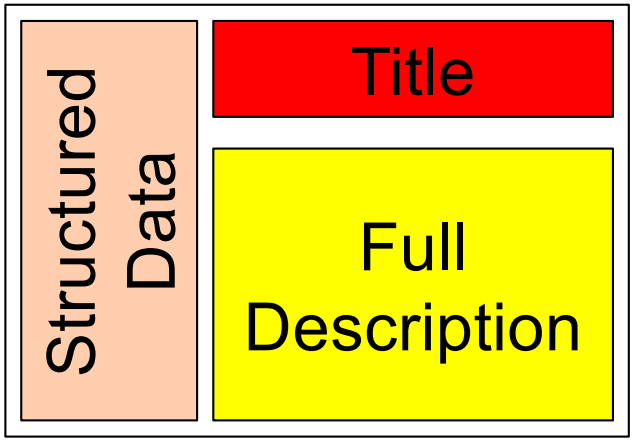
Applicants will have circa two years' experience in a credit control environment and be keen to develop a career in this area. The successful candidate will have confidence, be a good team player, and possess good IT skills

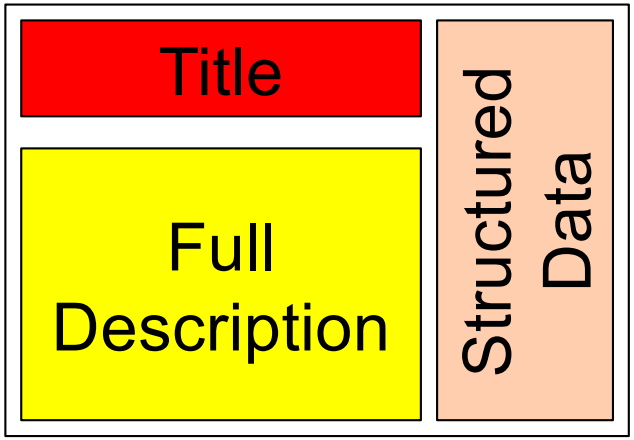
Information to extract:

- Title, Full Description (Free Text)
- Geographic area, Contract Type, Wages, Sector, Education (semi-structured data), …

Custom Information Extraction

- Each web site has its own layout
- Multiple layouts among same site pages
- Information extraction is performed ad hoc for each page structure





Duplicate Removal Overview

- **Duplicate Detection**: to identify
  - duplicate job offers posted on different Web source
  - job Vacancies published multiple times (on the same site)
- Two vacancies are considered duplicate if
  - Text similarity (edit distance < chosen threshold)
    - edit_distance(“Play”, ”Plays”)=1
  - Publishing date distance < 1 month

Text Classification

- Classifier: a function that maps (namely classifies) a data **item** into one of several **predefined classes**
  - **Items**: web job vacancies (title + full description)
  - **Predefined classes**: ISCO 4th digit classification
  - Methodology: exploit machine learning approach (supervised learning)
    - Algorithms trained over a set of (already labelled) documents
    - Once trained, the algorithm can classify unlabelled documents

Classification Accuracy

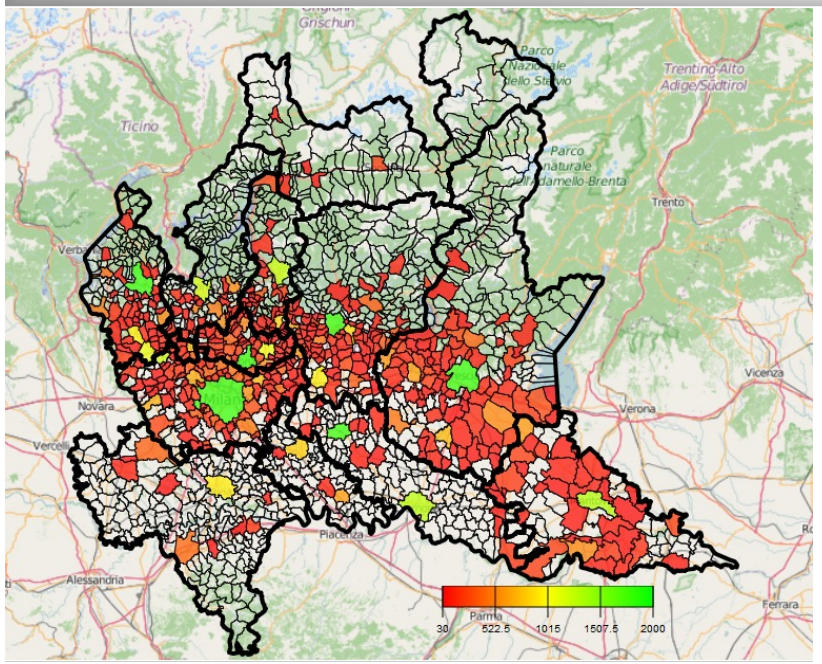
- Labelled data split into 2 datasets: training and test
- Classifier trained on the training set, then evaluated on the test set
- No vacancy used in training was reused during test
- Accuracy =
  (N. vacancies correctly classified) / (N. vacancies)
  - Focus on ISCO 4th digit codes
  - Vacancy correctly classified:
    all 4 digits correctly predicted
- What is the achieved accuracy, in your opinion?

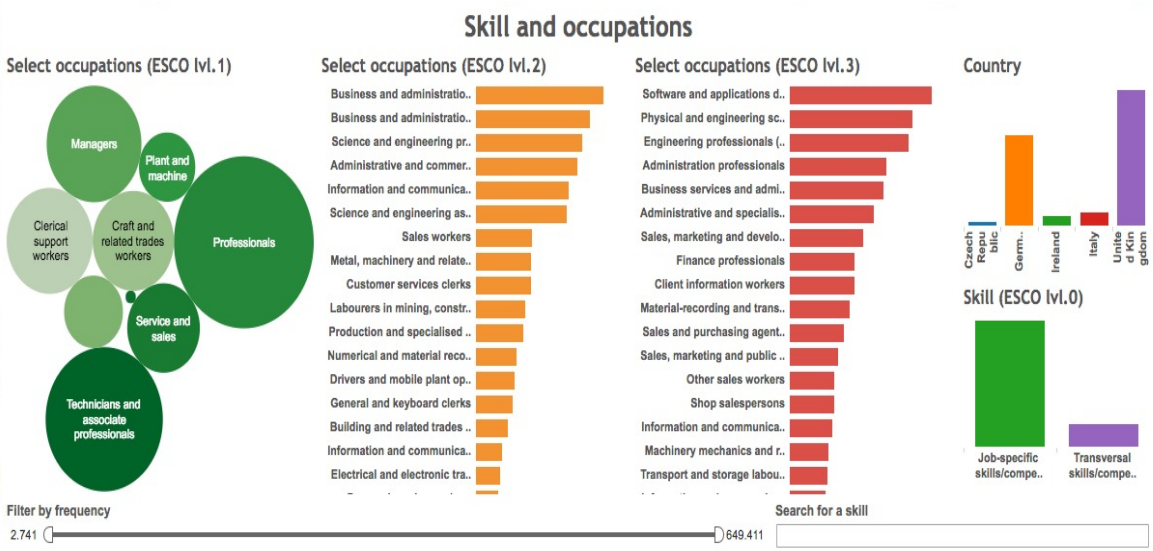
Language	Accuracy
CZ	98%
UK	80%
DE	79%
IT	80%

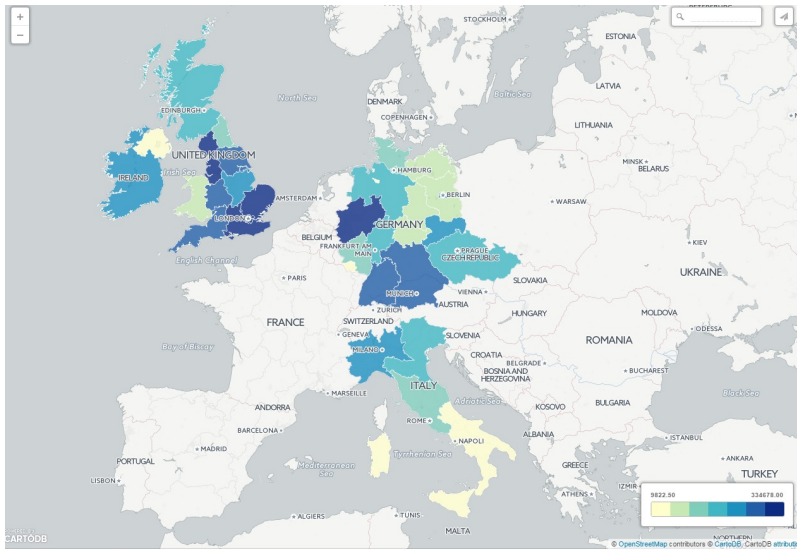
Accuracy Upper Bounds?

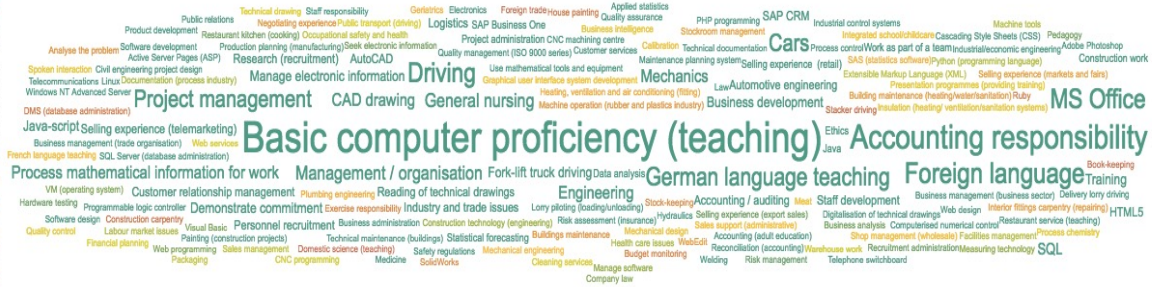
- The Czech classifier achieved a 98% accuracy
  - **Large set** of **high quality** 4th digit ISCO labelled vacancies available from public websites
  - Vacancies consistently labelled
- Other language classifiers
  - English classifier accuracy raised from 40% to 80% by improving the training set classifications …
  - Similarly the other languages

Analysis Examples









Q & A

Thank you for your attention

Are there any questions?

Selectedd references

- ovaglio, P., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. STATISTICAL ANALYSIS AND DATA MINING, 11(2), 78-91.
- Boselli R., Cesarini M., Marrara S., Mercorio F., Mezzanzanica M., Pasi G., and Viviani M.. Wolmis: a labor market intelligence system for classifying web job vacancies. Journal of Intelligent Information Systems, Sep 2017.
- Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data." Communications of the ACM 39.11 (1996): 27-34.
- Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." Ann Arbor MI 48113.2 (1994): 161-175.
- Cohen, Aaron M., and William R. Hersh. "A survey of current work in biomedical text mining." Briefings in bioinformatics 6.1 (2005): 57-71.
- Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." The Journal of Machine Learning Research 2 (2002): 45-66.
- Amato, Flora, et al. "Challenge: Processing web texts for classifying job offers." Semantic Computing (ICSC), 2015 IEEE International Conference on. IEEE, 2015.
- R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, “Hsow the social media contributes to the recruitment process,” in Proceedings of European Conference on Social Media (ECSM 2014), 2014, pp. 10–11.
- R. Boselli, M. Mezzanzanica, M. Cesarini, and F. Mercorio, “Planning meets data cleansing,” in The 24th International Conference on Automated Planning and Scheduling (ICAPS 2014). AAAI, 2014, pp. 439–443.
- R. Boselli, M. Mezzanzanica, M. Cesarini, and F. Mercorio, “A policy- based cleansing and integration framework for labour and helthcare data,”inKnowledgeDiscoveryandDataMining,LNCS8401. Springer, 2014, pp. 141–168.
- M. Mezzanzanica, R. Boselli, M. Cesarini, and F. Mercorio, “A model-based evaluation of data quality activities in KDD,” In- formation Processing and Management (in press, available at http://dx.doi.org/10.1016/j.ipm.2014.07.007), 2014.

Source Selection

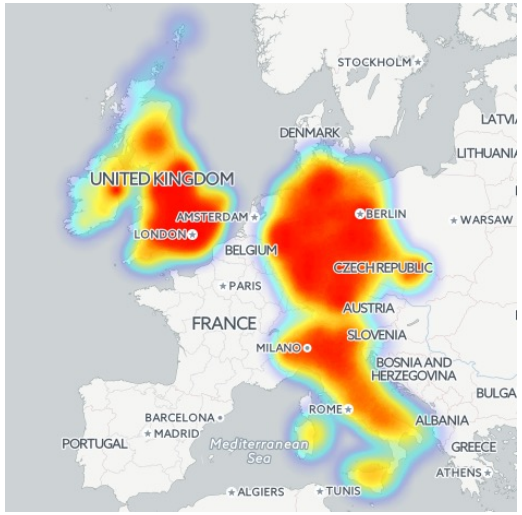
- Web site selection by a group of experts (few web sites per country scraped)
- Criteria (number of posts per month, fields/sectors covered …)

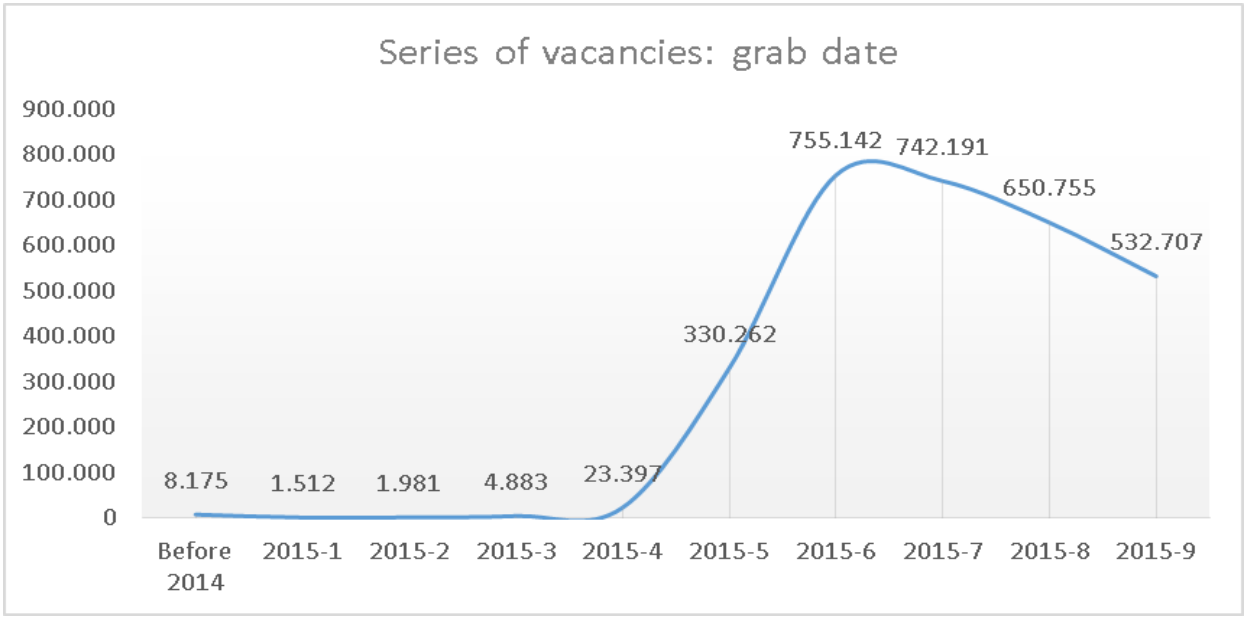
COUNTRY	SELECTED WEBSITES
CZECH REPUBLIC	• Annonce.cz • Monster.cz • Portal.mpsv.cz
GERMANY	• Gigajob.de • Online-stellenmarkt.net • Jobs.meinestadt.de
UK	• reed.co.uk • cv-library.co.uk • monster.co.uk • totaljobs.com
IRELAND	• Irishjobs.ie • Jobs.ie • Irishtimes.com/jobs
ITALY	• Infojobs.it • Monster.it • Adecco.it

Some numbers

3.051.005 job vacancies (duplicates removed)

- 1’754’064 UK
- 988’470 Germany
- 140’915 Italy;
- 91’522 Czech Republic
- 76’034 Ireland





Structured Data Reconciliation

- Structured/Semi-structured data
  - Geographic area (NUTS)
  - Contract Type
  - Working hours
  - Sector (NACE)
  - Education
  - …
- Processing
  - Site specific data mapped on standard taxonomies
  - Mapping developed for each language/site

Data	Taxonomy Entry
London	Inner London
Inner Lond. North	
Greenwich	
…	

Skill Extraction

- Word windows are extracted from full descriptions using sentinel words (or expressions)
- Skills are identified among the selected windows using a dictionary look-up approach
- A dictionary was developed for each language
- Dictionary creation
  - Data driven
  - (Man-in-the-loop) supervised incremental process