

Real Time Labour Market Information

Mario Mezzanzanica
Mirko Cesarini

mario.mezzanzanica@unimib.it
mirko.cesarini@unimib.it



CEDEFOP



4 Million Web Job Advertisements

Job Summary

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

- Each vacancy
 - Free Text
 - Semi structured data e.g.,
 - Location
 - Contract type

**HOW TO DEAL WITH
MILLIONS
OF WEB JOB VACANCIES**



Challenging Issues

- Collecting Job Advertisements from heterogeneous Web Sources
- Filtering and Extracting information from raw texts
- Classify Job Advertisements to the *proper* ISCO (4th digit) code / ESCO Occupation group
- Extracting skills
- Analyze and Visualize data for decision making purposes

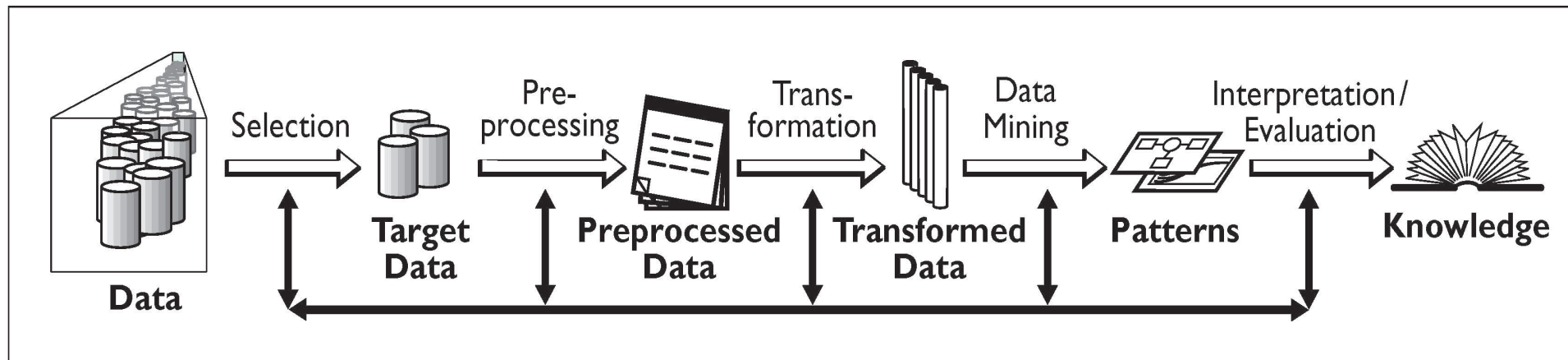


Terms - Settings

- **ESCO Occupations:** a *grouping* of jobs involving similar tasks and which require a similar skills set
- **Web Job vacancy:** a free text specifying a job position that will be classified on one ESCO Occupation
- **Web Skill:** a combination of words extracted, processed and filtered from web Job vacancies

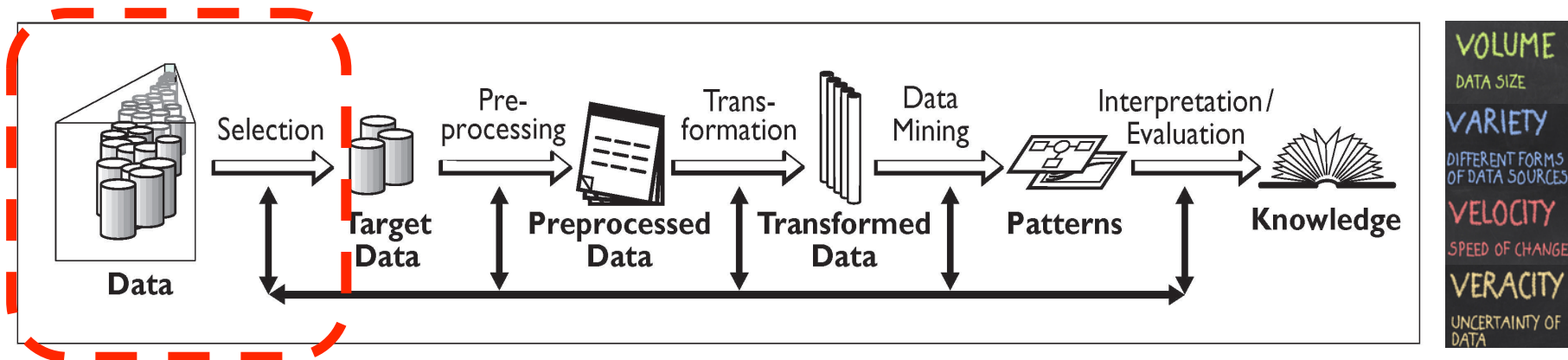
Knowledge Discovery in Databases

Turning data into knowledge (Fayyad, 1996)





Step 1: Sources Selection and Scrapping



Each Web source was evaluated and ranked on the basis of the following criteria:

- Typology
- Size
- International
- Industry
- Publication date
- Update
- Type
- Quality of Descriptions
- Number of structured variables
- Download limit



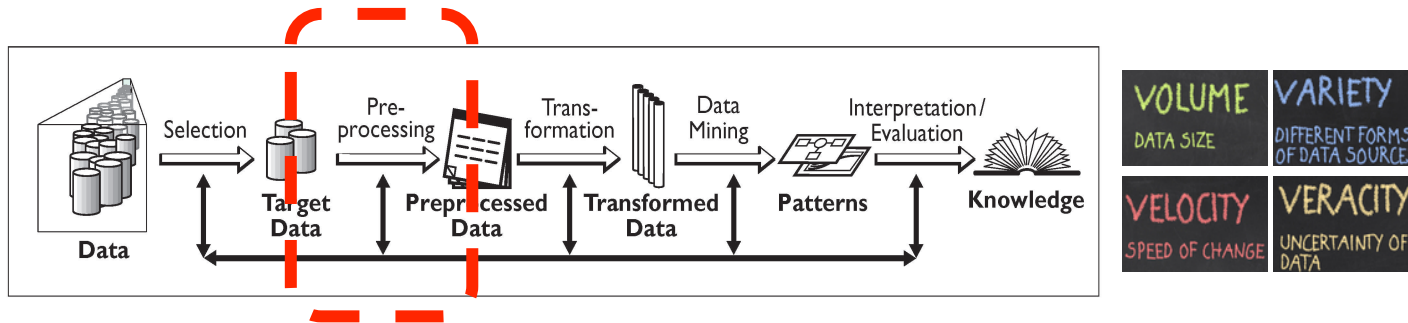
Country	Web Site
CZ	monster.cz
	portal.mpsv.cz
DE	Gigajob.de
	online-stellenmarkt.net
	jobs.meinestadt.de
UK	reed.co.uk
	cv-library.co.uk
	monster.co.uk
	totaljobs.com
IE	irishjobs.ie
	jobs.ie
	irishtimes.com/jobs
IT	infojobs
	monster.it
	adecco

Websites & Scraped Data

Language	# Downloaded Vacancies
IT	276,909
CZ	100,055
DE	1,555,924
IE	152,661
UK	2,142,942
Total	4,228,491



Step2: Data Deduplication



- **Duplicate Detection:** identify
 - duplicate job offers posted on different Web source
 - job Vacancies published multiple times (on the same site)
- Some numbers:
 - Downloaded Vacancies 4'228'491 (100.00%)
 - Vacancies after duplicate removal 2'980'546 (70.49%)



Step 3: Extraction

Job Summary Credit Control Assistant	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
Location London	
Industries Accounting and Auditing Services	
Job Type Full Time Permanent	

We currently have a vacancy for a Chef / Cook to work with our client's busy team at their Restaurant within their luxury home furnishing store. As Chef / Cook the successful candidate must have good previous experience and preferably hold a current Health & Hygiene Certificate. Duties will include preparing and cooking fresh food to order plus other essential kitchen duties. As Chef / Cook you'll need to demonstrate: that food is prepared to the highest standards in accordance with company specifications, paying particular attention to food hygiene; Food rotation is maintained to minimise wastage; An ability to control stock levels keeping within restaurant budgets; the ability to work on own initiative and as part of an effective team.			
Job Summary Chef	Location Dublin	Industries Food preparation	Job Type FT – Open Ended

■ Text extraction

■ Structured data

○ Extraction

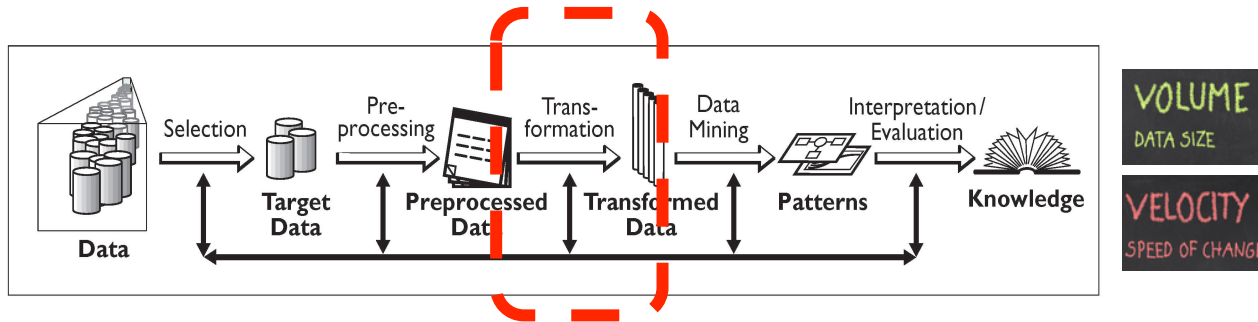
○ Reconciliation

• Full Time vs FT

• Permanent vs Open Ended



Step 4: Text Classification



- Classifier: a function that maps (namely classifies) a data **item** into one of several **predefined classes**
 - **Items**: web job vacancies
 - **Predefined classes**: ISCO classification system
- Methodology: exploit supervised learning approaches



Training.... Testing....



Job
Vacancies
Training
Set

Machine Learning (ISCO 1212)
2230)

Test

Staff Manager (ISCO 1212)



Machine
Learning
Algorithm

ISCO classes (4 digits code)

- > [Armed forces occupations](#)
- > [Clerical support workers](#)
- > [Craft and related trades workers](#)
- > [Elementary occupations](#)
- > [Managers](#)
- > [Plant and machine operators and assemblers](#)
- > [Professionals](#)
- > [Service and sales workers](#)
- > [Skilled agricultural, forestry and fishery workers](#)
- > [Technicians and associate professionals](#)



Classification Accuracy

- Accuracy: $\frac{\text{\#correctly classified}}{\text{\#vacancies}}$
 - Focus on 4 digit codes ...
 - ... vacancy matched on ISCO code: Yes or No
- Classification accuracy computed against a test set:
 - Web job vacancies already classified (validated by experts)
 - Machine Learning output compared with the (already known) result
- Training and test set built by domain experts

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

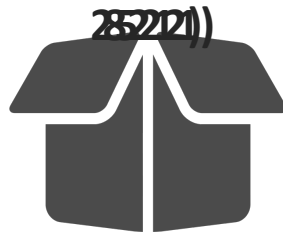


Classification Example: Classify Web Job Vacancies



Job Vacancies
to be
classified

Classified as
Subordinate (5105(1910)
25221))



Machine
Learning
Algorithm

ISCO classes (4 digits code)

- > [Armed forces occupations](#)
- > [Clerical support workers](#)
- > [Craft and related trades workers](#)
- > [Elementary occupations](#)
- > [Managers](#)
- > [Plant and machine operators and assemblers](#)
- > [Professionals](#)
- > [Service and sales workers](#)
- > [Skilled agricultural, forestry and fishery workers](#)
- > [Technicians and associate professionals](#)



4Digits ISCO Classification

- I am seeking a C# & VB.NET Developer / Analyst Programmer to join a skilled team which is tasked with building a range of new systems for one of the UKs best known companies based in Watford, Hertfordshire (close to Watford Junction and excellent transport links!).
 - As a C# & VB.NET Developer / Analyst Programmer you will get the chance to work on globally critical web systems and products for a household name. Within the role you will code and perform systems analysis as well as reviewing and providing expertise and advice for the progression and update of the global web based systems.
 - Skills required for the C# & VB.NET Developer / Analyst Programmer: Commercial experience of C# and VB.NET, Broad technical knowledge including XML and SQL Server would be ideal, The ability to work on the end to end software development lifecycle
 - Experience of liaising with different business departments, Any experience of retail or EPOS systems would be ideal but by no means essential as full training will be provided.
-
- How to classify the advertisement?
 - 2511 Systems analysts
 - 2512 Software developers
 - 2513 Web and multimedia developers
 - Sometimes misclassification is not an error!



4Digits ISCO Classification

- **Credit Control Assistant.** 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
-
- How to classify the advertisement?
 - 2411 Accountants
 - 3313 Accounting associate professionals
 - Sometimes misclassification is not an error!



Occupation Families based on Skill Similarities

Family groups of ESCO occupations sharing *at least* 50% ESCO skills each other (within the same family group)

Occupation set

...

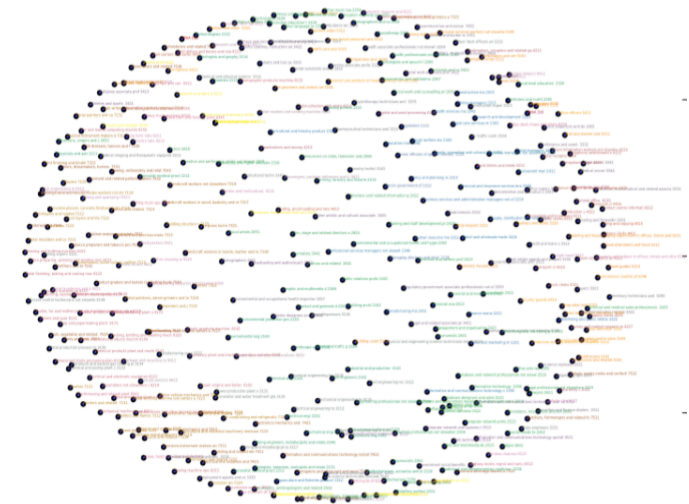
Field crop and vegetable growers, Tree and shrub crop growers, Mixed crop growers

Power production plant operators, Incinerator and water treatment plant operators

Chefs, Cooks

Building frame and related trades workers not elsewhere classified, Civil engineering labourers

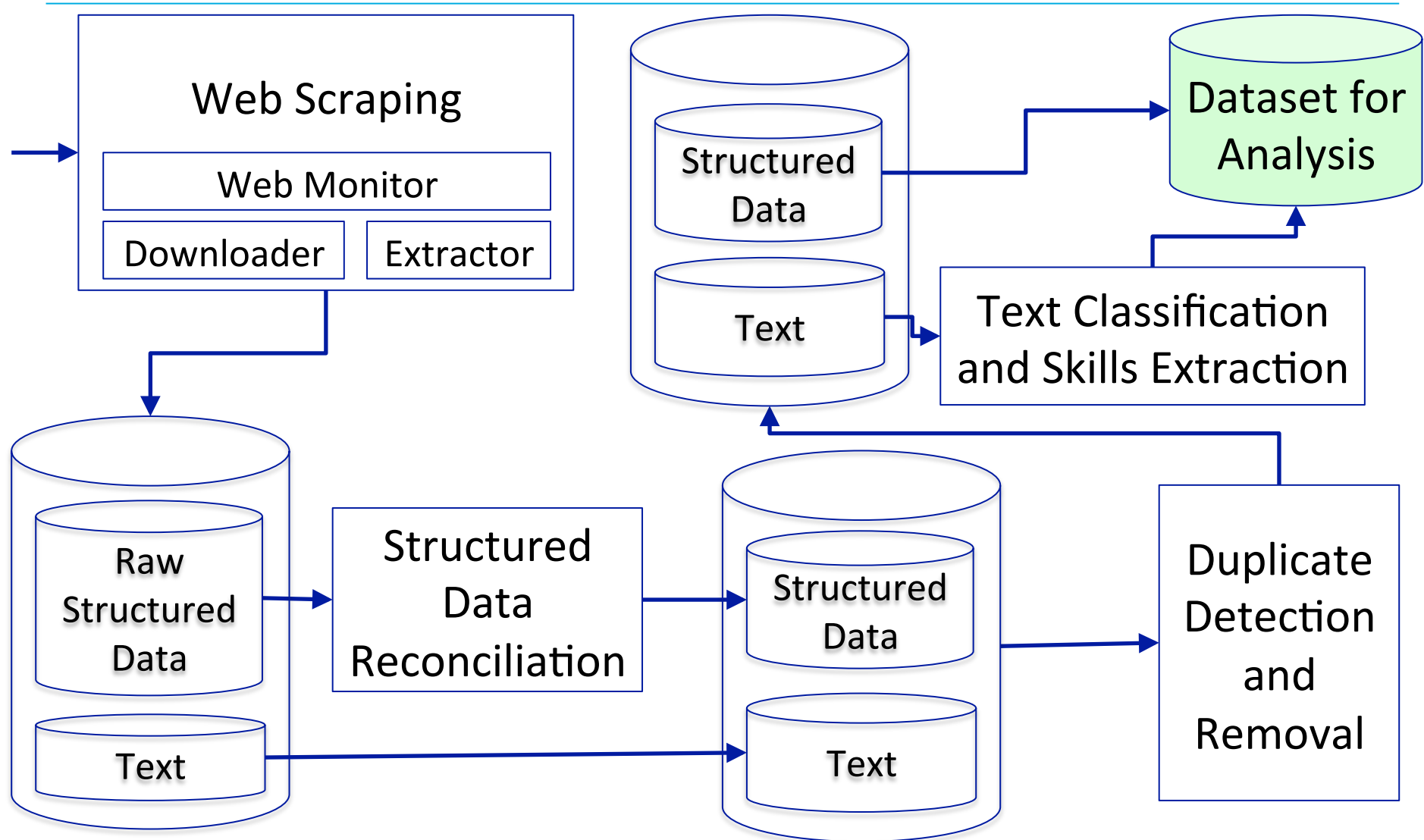
...



- 2-dimensional representation of occupations
- Occupations closeness driven by shared skills

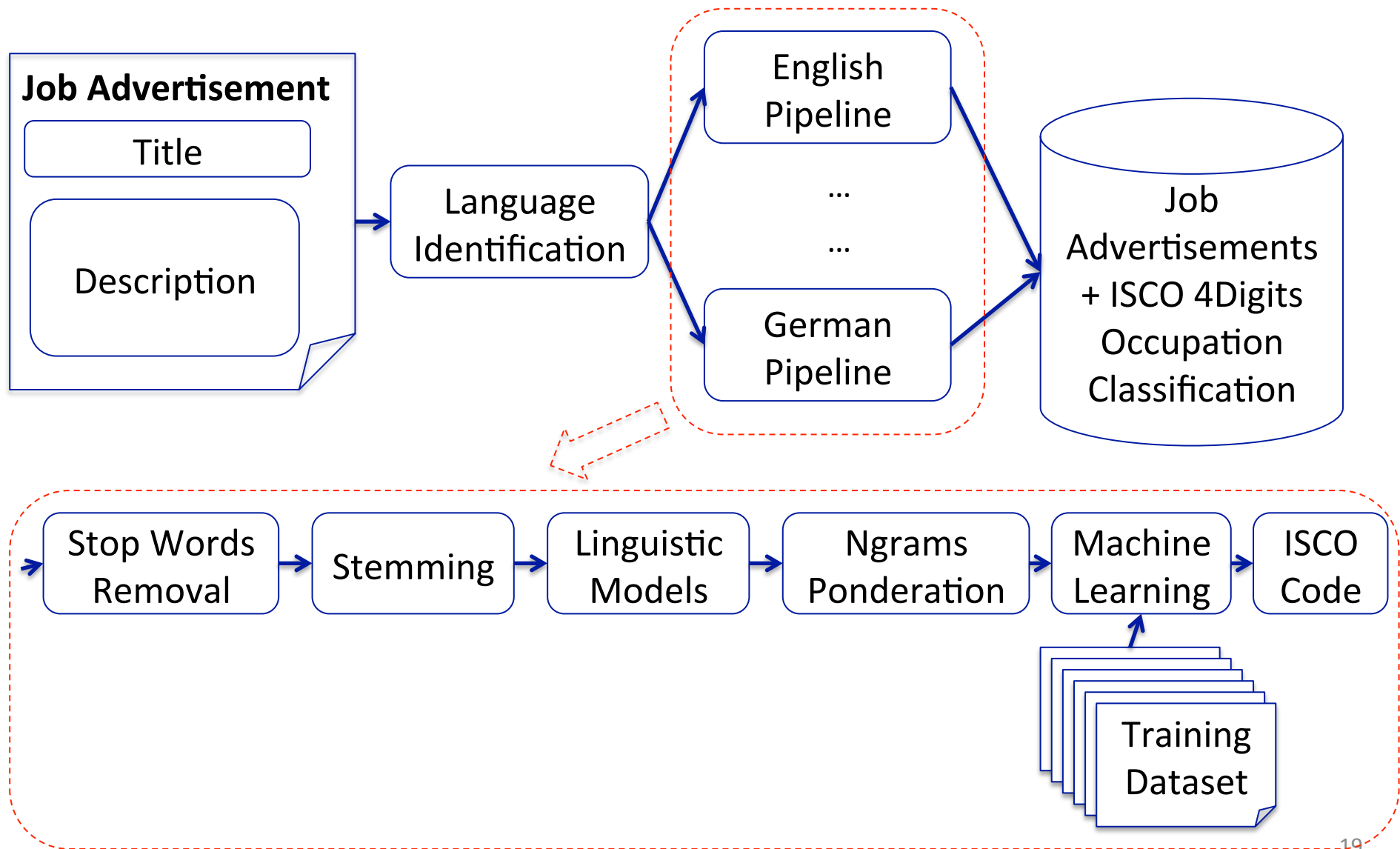


Overall Data Flow





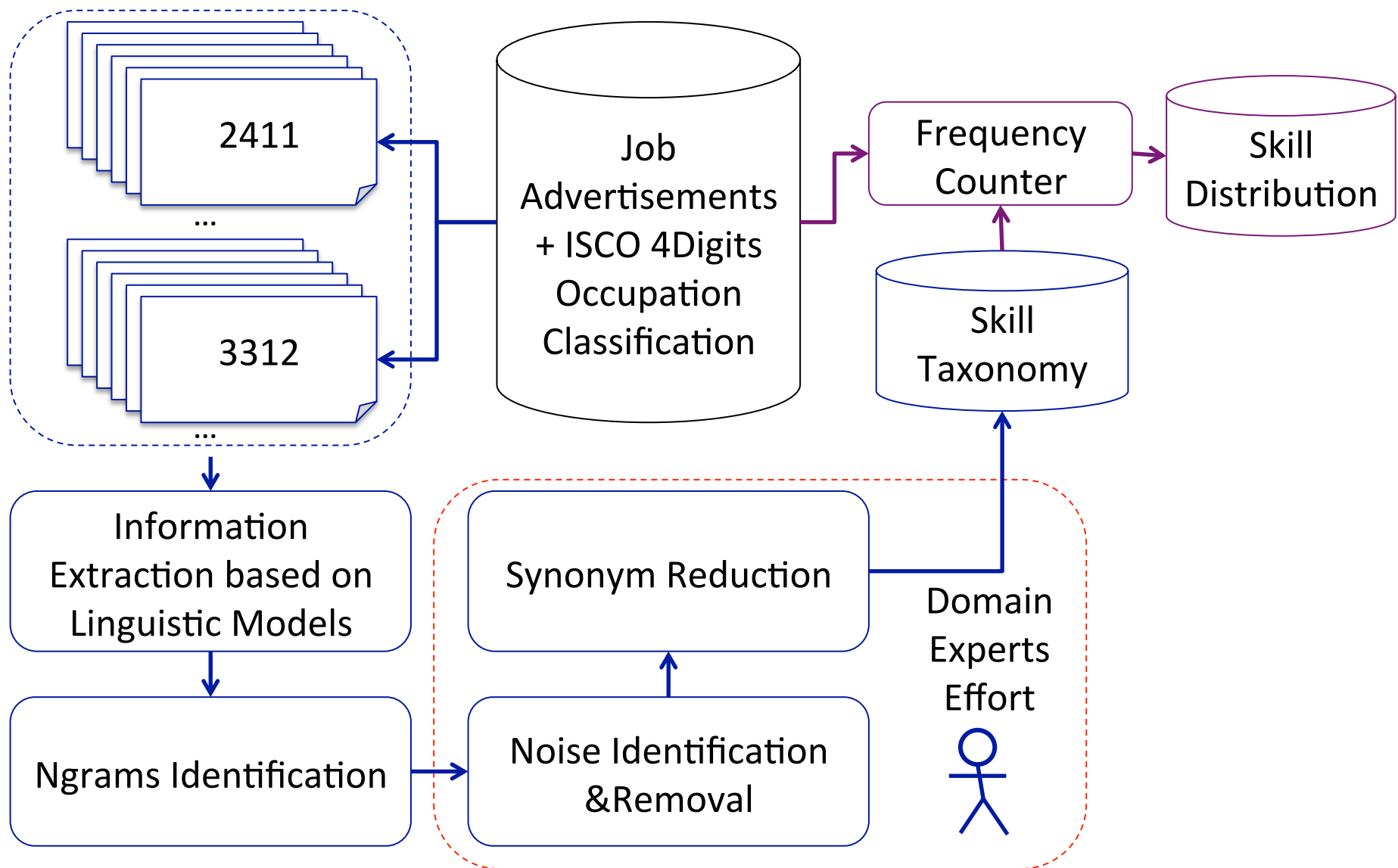
Focus on Text Classification





Web Skills Extraction

(Focus on a single Language Pipe)





Web Skill Taxonomy Building

- Taxonomy building: semi-automated process
 - Computer supported process
 - Domain Experts helps identifying:
 - synonyms
 - frequency distributions to identify noise (tuning)
- Iterative process, man-in-the-loop
- Further human effort
 - Optional e.g., once per year
 - Related to lexicon evolution
 - Incremental process: every step benefits/exploits the previous one results
 - synonyms sets don't change completely in a year!
 - noise distribution may be regular over time



Web Skills Example

Credit Control Assistant. Our client is a well-known group of companies specialising in import and wholesale of consumer 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**



Q & A

- Thank you for your attention
- Are there any questions?

Selected references

- 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, "How 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 healthcare data," in Knowledge Discovery and Data Mining, LNCS 8401. Springer, 2014, pp. 141–168.
- M. Mezzanzanica, R. Boselli, M. Cesarini, and F. Mercorio, "A model-based evaluation of data quality activities in KDD," Information Processing and Management (in press, available at <http://dx.doi.org/10.1016/j.ipm.2014.07.007>), 2014.