Applying machine learning tools on web vacancies for labour market and skill analysis

Emilio Colombo[1] Fabio Mercorio[2] Mario Mezzanzanica[3]

[1]Catholic University of Milan and Crisp [2]University Milano-Bicocca and Crisp
[3]University Milano-Bicocca and Crisp

Feb 2018

New tools for analyzing the labour market of the future

- Technology and globalization are radically changing labor markets throughout the world
- Some jobs are disappearing, new jobs are emerging, existing ones are changing.
- In particular jobs' skill requirements is changing considerably.
- Which occupations will grow in the future and where?
- What skills will be needed in the future?
- These are challenging questions that need new tools and new data

Why web vacancies can help

- We need tools able to address the complexity of labor market developments
- We need to focus on skills which are susceptible to change
- We need a data driven approach
- Web vacancies: bottom up approach, real time data, entirely data driven, cheap, very rich
- Surveys (e.g. skill survey): top down approach, pre-defined, expensive, low frequency data

Web vacancies challenges

- Totally unstructured data
- Need classification
- Lots of data, duplication issues
- Representativeness

The Wollybi project

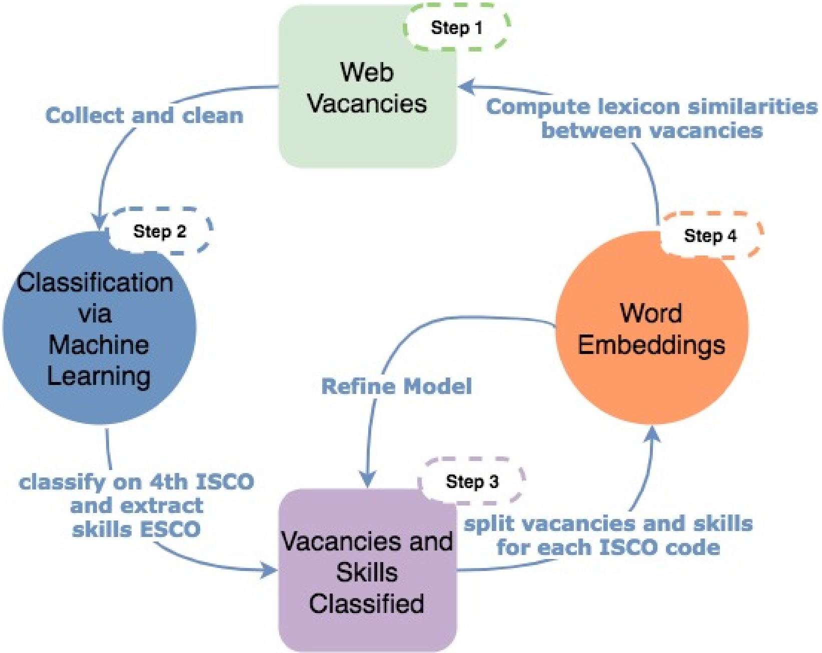
- Since 2013 collects and analyzes web vacancies Italy
- 2.8 millions unique obs analyzed
- Now in charge of a European project funded by Cedefop to implement a European system of LMI.
- By 2018 the project will analyse data from Italy, Germany, France, Spain, UK, Ireland, Czech Rep. (2/3 of EU pop.)
- By 2020 the whole EU will be analysed
- Lots of challenges in a multi language setting!

Wollybi data

- Vacancies
- Job title (ESCO/ISCO)
- Skills (ESCO)
- Additional skills (job requirements)
- Georeferencing (Nuts 4, town)
- Industry
- Education (work experience)

Methods and tools

- Big data tools, (source selection, de-duplication, data cleaning) textual analysis
- Machine learning tools for the classification process
- Skills extraction, n-grams.
- How skill requirement evolve over time? Word embedding (word2vec, doc2vec) to compute index of similarities between skills and vacancies.
- How detect new emerging occupations? LDA


Step 1
Web Vacancies
Collect and clean
Compute lexicon similarities between vacancies
Step 2
Classification via Machine Learning
Step 4
Word Embeddings
Refine Model
classify on 4th ISCO and extract skills ESCO
Step 3
Vacancies and Skills Classified
split vacancies and skills for each ISCO code

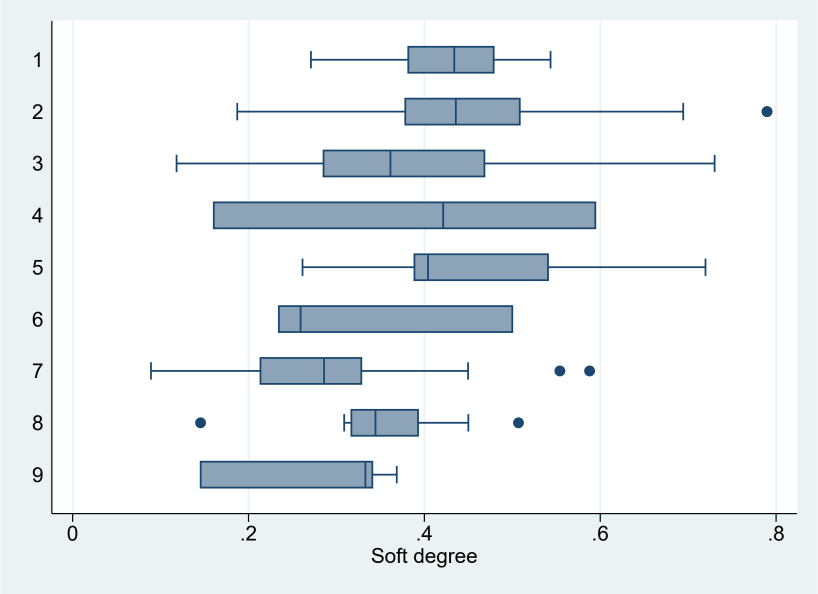
Hard and soft skills

- **Hard skills**: typically job-specific skills and competences that are needed to perform a specific job or task (e.g. knowledge of specific software or instruments, specific manual abilities etc.)
    - *Digital skills* are a sub-class of hard skills
- **Soft skills**: transversal competences which refer to the capacity of individuals to interact with others and the environment (e.g. communication skills, problem solving etc.).

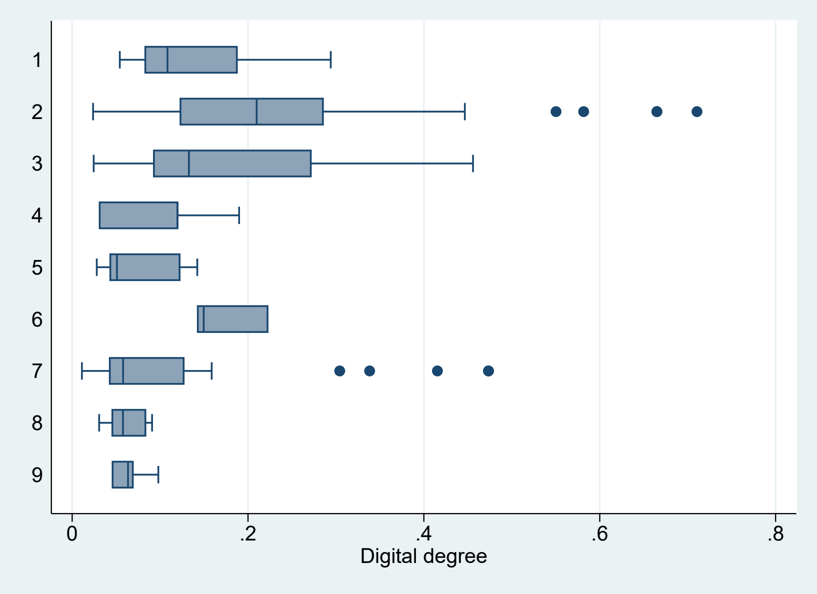
Digital skills

- **Information Brokerage skills**. Ability to use ICT tools and platforms for data exchange and communication (e.g. social media);
- **Basic ICT skills**. Ability to use some ICT specific applications for supporting standard individual professional activities (e.g. Office suite);
- **Applied/Management ICT skills**. Tools and software used within the organisation for supporting management, operational and decision making processes (e.g. administrative software);
- **ICT Technical skills**. Solutions, platforms and programming languages that are ICT-specific (e.g. programming languages, advanced ICT software).

Soft skills distribution



ICT skills distribution



Explaining the probability of automation

	OLS	W.OLS	OLS	W.OLS
Soft skills	-0.727***	-0.645***		
	(0.122)	(0.001)		
Hard skills			0.842***	0.714***
			(0.118)	(0.001)
Digital skills			-0.719***	-0.683***
			(0.110)	(0.001)
Const.	0.817***	0.759***	0.169**	0.224***
	(0.045)	(0.000)	(0.079)	(0.001)
R2	0.065	0.058	0.137	0.148
N.	512	512	512	512

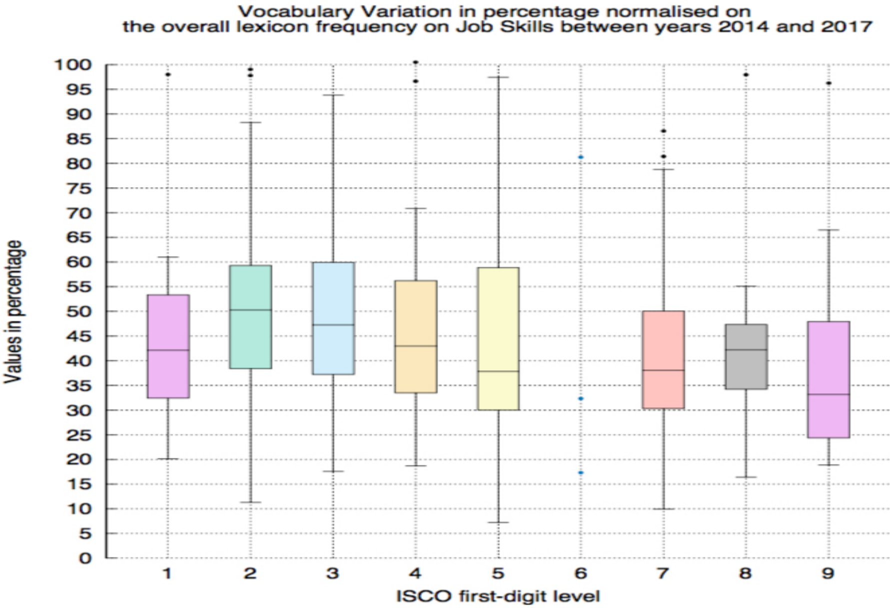
Explaining the probability of automation

	Q12 Prob	Q34 Prob	Q12 Prob	Q34 Prob
Hard s.	0.589***	-0.008	0.573***	0.016
	(0.106)	(0.048)	(0.105)	(0.049)
Digital s.	-0.493***	0.101**		
	(0.099)	(0.044)		
Inf. Brok.			-0.048	0.015
			(0.381)	(0.148)
ICT Technical			-0.694***	-0.190*
			(0.179)	(0.106)
Basic ICT			-1.264***	0.383***
			(0.349)	(0.139)
App. Man. ICT			-0.249	0.052
			(0.183)	(0.093)
Const.	-0.004	0.868***	0.025	0.853***
	(0.064)	(0.034)	(0.067)	(0.036)
R2	0.143	0.021	0.182	0.068
N.	256	256	251	249

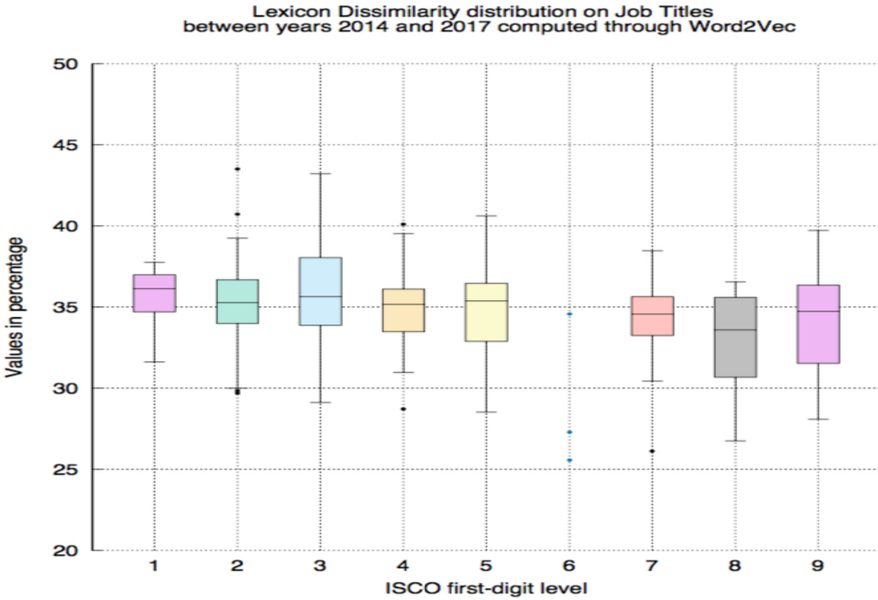
How does the skill content of occupation change?

- We have measured the change in description and wording within occupations (4 digit ISCO) between 2014 and 2017
- We have measured the change in vocabulary used for titles and skills
- We have measured the change in *lexicon* used for titles and skills applying the Doc2Vec technique

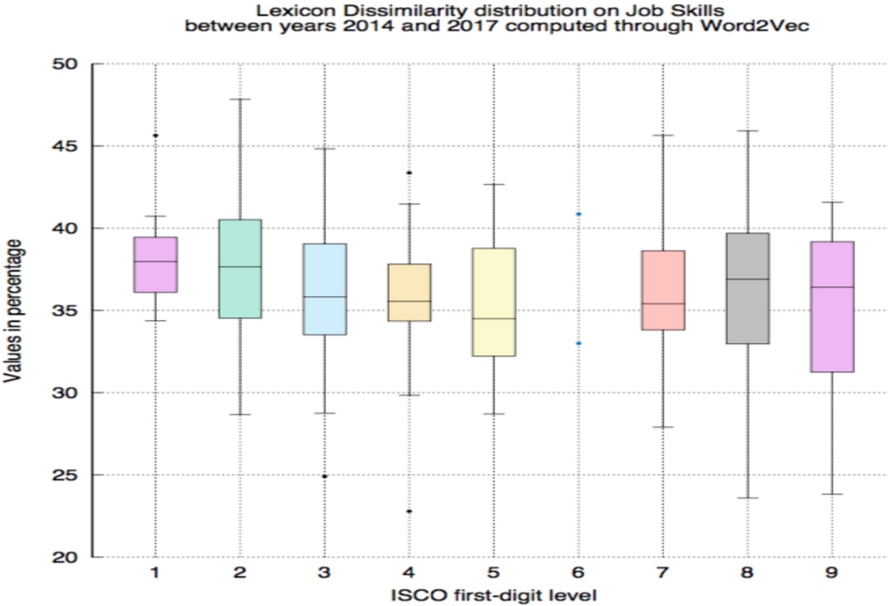
The change in vacancy content: skill vocabulary



The change in vacancy content: title lexicon



The change in vacancy content: skill lexicon



How to detect new emerging occupations?

- We take vacancy titles by ISCO 4D. We apply LDA (Latent Dirichlet Allocation) which returns returns a list of topics.
- The topics are basically different clustering of words used in vacancy titles with a word probabilistic model.
- Topics are validated by experts which identify new occupations
- The skill set is applied to new occupations

New emerging occupations

New Occ.	**Hard Skill**	**Hard Skill**	**Hard Skill**
BI Analyst	*ICT (50.32%)*	*Maths&Stats (29.60%)*	*Bus.&Admin. (19.33*
	SQL & Oracle	Data Analysis	Public Relations
	SharePoint		Management
	Data Integ*		Manage Quality?
Data Scientist	*Maths&Stats (48.15%)*	*ICT (29.01%)*	*Bus.&Admin.(22.84%*
	Data Analysis	SQL & Java	P.R.
	SAS	Business Intell.	Management
	SAP & SPSS*	Data Integration*	Customer Rel. Manag
Facility Man.	*Bus&Admin.(48.07%)*	*ICT (19.89%)*	*Law (16.02%)*
	Management	MS Office	Security Law
	Public Relations	AutoCAD	Legal Studies
	Negotiation exp	Basics ICT	
HSE Spec.	*Law (41.51%)*	*ICT (28.30%)*	*Bus.&Admin. (16.98*
	Security Law	Basic ICT	Manage Quality
	Legal Studies	SAP CRM	Industry Systems

New emerging occupations

New Occupation	Soft Skill	Soft Skill	Soft Skill	Soft Skill
BI Analyst	Foreign languages (39.23%)	Positive attitude (31.51%)	Problem solving (15.35%)	Cooperation with others (6.27%)
Data Scientist	Positive attitude (38.58%)	Foreign languages (29.63%)	Problem solving (13.58%)	Cooperation with others (6.79%)
Facility Manager	Foreign languages (39.68%)	Positive attitude (28.95%)	Leadership ability (16.09%)	Problem solving (11.53%)
HSE Specialist	Foreign languages (14.55%)	Problem solving (56.36%)	Positive attitude (20.00%)	Cooperation with others (3.64%)