The STEM requirements of “Non-STEM” jobs: Evidence from UK online vacancy postings

Inna Grinis

ILO Workshop on big data for skills anticipation and matching

19-20 September 2019

Motivation

- **The UK spends more money on STEM (Science, Technology, Engineering, Maths) education than on non-STEM one …**

- STEM in the 2017's spring Budget: "support for 1,000 PhD places, *particularly for those studying STEM subjects*''

- STEM education more heavily subsidized by the HEFCE – most STEM disciplines "high-cost" and "strategically important", whereas most non-STEM ones classified as "classroom-based"

- **… but less than half of STEM graduates work in "STEM" occupations** (e.g. *Scientists, Engineers*)

"STEM pipeline leakage''

problematic if "non-STEM" recruiters do NOT require and value STEM knowledge and skills because:

- wastage of resources
- creates shortages in STEM occupations

Question

To what extent do recruiters in "non-STEM" occupations require and value STEM knowledge and skills?

- The UK economy is hit by trends like **digitization**, the arrival of **Big Data**…

"A whole range of STEM skills - from statistics to software development - have become essential for jobs that never would have been considered STEM positions. Yet, at least as our education system is currently structured, students often only acquire these skills within a STEM track."

Matthew Sigelman (CEO of Burning Glass Technologies)

- Examples of keywords from online vacancy postings of:

Graphic designers: *"JavaScript", "HTML5", "User Interface (UI) Design", "jQuery", "Computer Software Industry Experience", "Computer Aided Draughting/Design (CAD)"…*

Management consultants and business analysts: *"SQL", "Data Warehousing", "Optimisation", "Data Mining", "Microsoft C#", "Relational Databases", "Big Data" …*

Artists: *"Python", "Auto CAD", "3D Modelling", "3D Design", "Autodesk", "Microsoft C#", "3D Animation", "Computer Software Industry Experience" …*

Main Contribution & Results

- Existing studies identify STEM jobs at the **occupation-level**: *UKCES (2013, 2015), DIUS (2009), BIS (2011), Mason (2012), Rothwell (2013)…*

STEM occupations

Identified using judgment, % STEM degree holders, O*NET Knowledge scales …

STEM jobs

Jobs belonging to STEM occupations

- New **job-level** approach using UK online vacancies data and machine learning classification algorithms:

STEM disciplines

Sciences, Technology, Engineering, Mathematics

STEM keywords

"Systems Engineering ", "3D Modelling", "C++"…

STEM jobs

Pr(STEM graduate|Keywords) > Pr(Non-STEM|Keywords)

- **Main results**: 1. STEM jobs ≠ STEM occupations

  35% of STEM jobs belong to non-STEM occupations

  15% of all vacancies in non-STEM occupations correspond to STEM jobs

  2. STEM jobs in non-STEM occupations are associated with higher wages

-> A significant proportion of non-STEM recruiters do require and value STEM knowledge and skills

Outline

1. **Data**
2. **Identifying STEM keywords & jobs**
3. **STEM jobs in the UK**

   Occupational & Spatial distributions

   The wage premium for STEM

   The STEM requirements of “Non-STEM” jobs

Data

- **Burning Glass Technologies (BGT):**
  - leading US labour market analytics company
  - collect, deduplicate and process online job ads
  - visit ~ 5,000 websites in the UK on a daily basis
  - 33 million job ads between Jan 2012- July 2016
- **Collect** job titles, locations, education, wages, etc. but also the **keywords from vacancy descriptions:**
  - taxonomy of 11,182 different keywords
  - any keyword with a match is picked up
  - taxonomy expands over time, historical re-parsing
  - vacancy description appears as a set of out-of-context keywords, e.g.:

"SAS – Writing – Data Collection – Econometrics – Project Design – Team Building – SQL - R"

  - ≥1 keyword for 90% of vacancies (median of 4-5)
  - median keyword appears in 173 (0.001%) postings

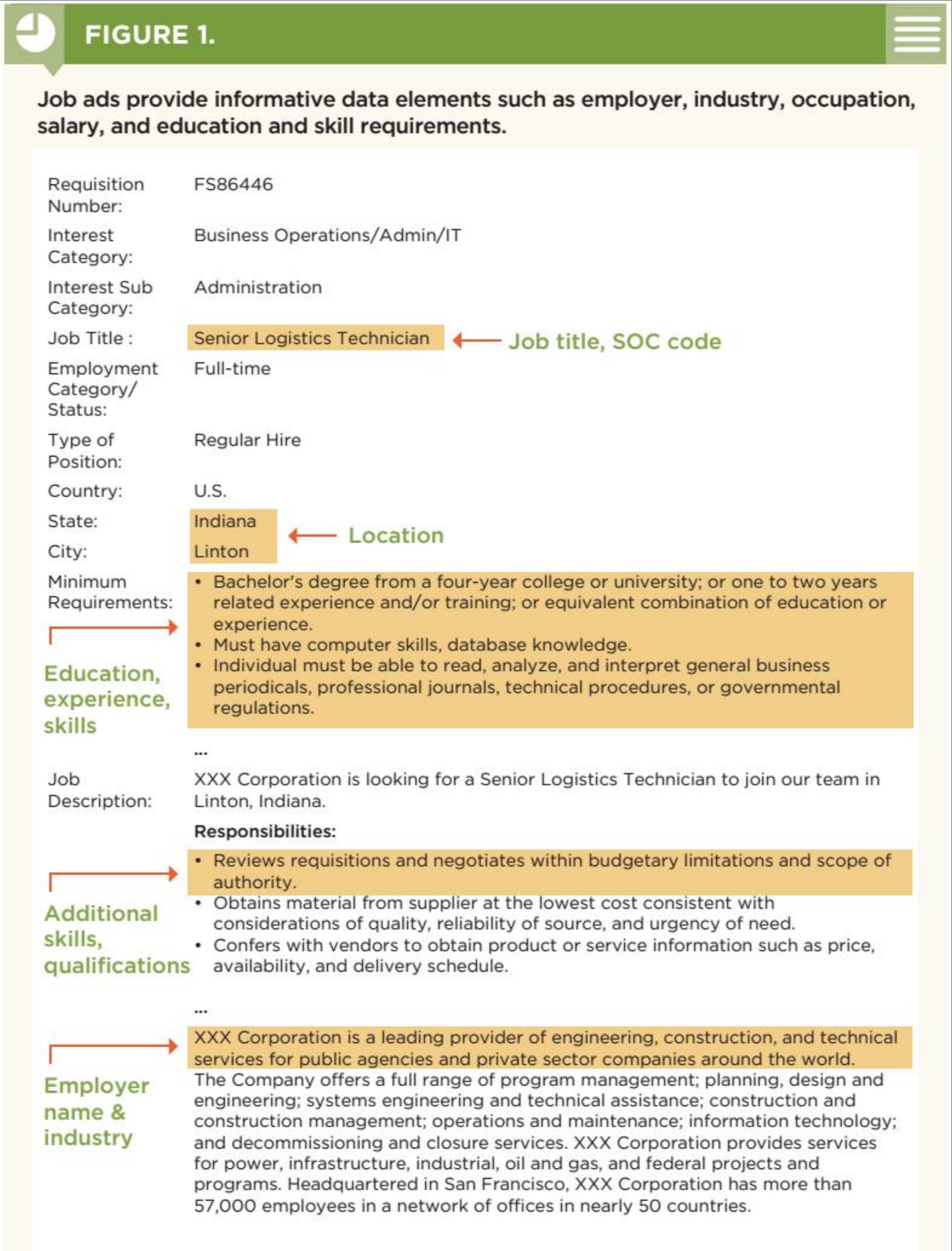


FIGURE 1.

Job ads provide informative data elements such as employer, industry, occupation, salary, and education and skill requirements.

Requisition Number:	FS86446
Interest Category:	Business Operations/Admin/IT
Interest Sub Category:	Administration
Job Title :	Senior Logistics Technician
Employment Category/ Status:	Full-time
Type of Position:	Regular Hire
Country:	U.S.
State:	Indiana
City:	Linton
Minimum Requirements:	• Bachelor's degree from a four-year college or university; or one to two years related experience and/or training; or equivalent combination of education or experience. • Must have computer skills, database knowledge. • Individual must be able to read, analyze, and interpret general business periodicals, professional journals, technical procedures, or governmental regulations. ...
Job Description:	XXX Corporation is looking for a Senior Logistics Technician to join our team in Linton, Indiana. **Responsibilities:** • Reviews requisitions and negotiates within budgetary limitations and scope of authority. • Obtains material from supplier at the lowest cost consistent with considerations of quality, reliability of source, and urgency of need. • Confers with vendors to obtain product or service information such as price, availability, and delivery schedule. ... XXX Corporation is a leading provider of engineering, construction, and technical services for public agencies and private sector companies around the world. The Company offers a full range of program management; planning, design and engineering; systems engineering and technical assistance; construction and construction management; operations and maintenance; information technology; and decommissioning and closure services. XXX Corporation provides services for power, infrastructure, industrial, oil and gas, and federal projects and programs. Headquartered in San Francisco, XXX Corporation has more than 57,000 employees in a network of offices in nearly 50 countries.

Source: Carnevale et al. (2014)

Data

- **Quality of the data**:

  - misclassifications, not all vacancies posted online (especially low-skilled), vacancies ≠ jobs

  + high correlations with official UK employment data (ONS): e.g. 0.94 for major occupations in 2014

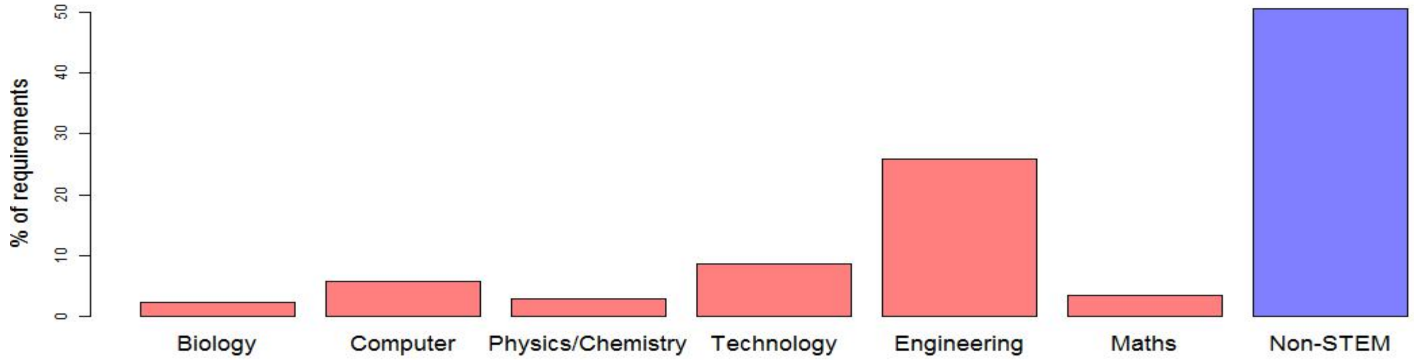
  - many missing values: e.g. occupation (0.5% missing), employer (69%), education (81%), wages (40%)…

- 12% of all vacancies contain **explicit discipline requirements,** e.g. *"Economics", "Chemistry"…*

  - high correlations with all vacancies: 0.94 (keyword frequency), 0.81 (4-digit occupations), 0.99 (counties)

  - 9566 different keywords (85.5% of taxonomy)

  - 425 distinct majors posted -> regroup into STEM and non-STEM disciplines

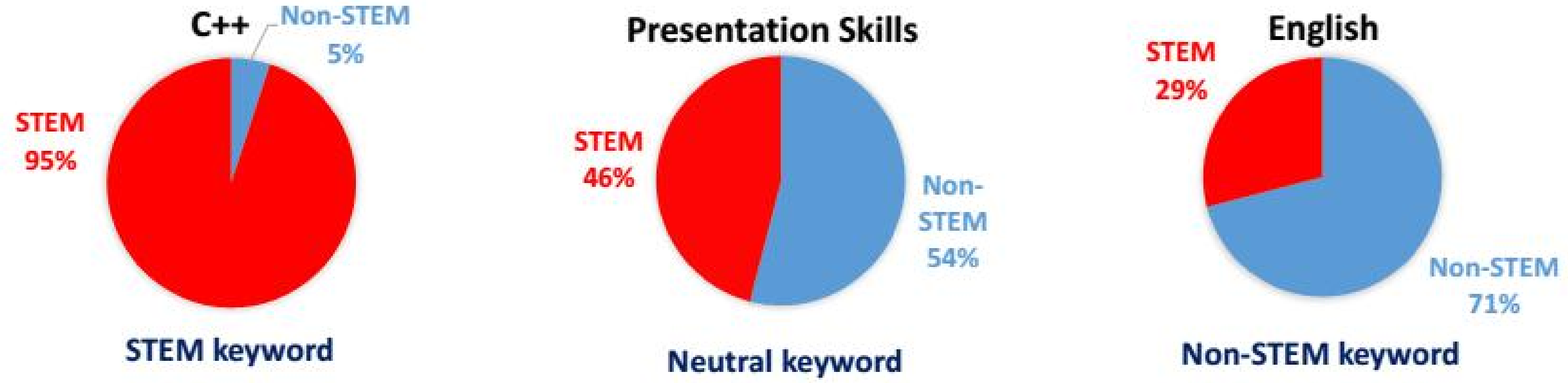


Note: Distribution of discipline requirements in the sample of 3.97m vacancies collected in Jan. 2012-Jul. 2016

Classifying Keywords

- **Objective**: classify 11k keywords into STEM and non-STEM
- **Challenge**: thousands of technical terms taken out of context, e.g.:

  "Leachate Management", "Actinic", "Step 7 PLC", "NASH", "Antifungal", "DFDSS"...
- **Solution:** design a systematic classification method
- **Strategy**: classify keywords depending on the discipline "contexts" in which they appear
- **Intuition**: A proper STEM skill, knowledge, task should rarely appear together with a non-STEM degree because it requires a proper STEM education and a STEM qualification, and vice versa



- **Main steps of the "context mapping" algorithm (unsupervised learning)**:

1. Record the distribution of disciplines with which a keyword appears
2. Implement K-means clustering on the distribution vectors to separate the keywords into STEM, Neutral, and Non-STEM
3. K-means clustering of STEM keywords into STEM domains

Classifying Keywords: Examples

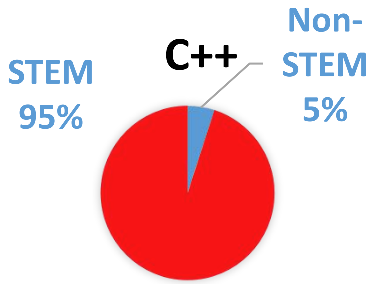
Computer Sciences keywords

Non-STEM keywords

Note: Random samples of around 100 keywords coloured and weighted by frequency of being posted.

Keyword "Steminess"

- **Steminess** of keyword *k* = % STEM discipline requirements with which k appears



$$steminess_{C++} = 0.95$$

Clusters	STEM	Neutral	Non-STEM
Median steminess	0.91	0.50	0.08
Mean steminess	0.89	0.49	0.10
Min steminess	0.69	0.29	0.00

- Steminess of *k* is the Maximum Likelihood estimate of $\Pr(STEM|k)$

From Keywords to Jobs: Multinomial Naive Bayes classifier

- Job j = **set** of keywords $K_j = \{k_1, k_2, \ldots, k_{n_j}\}$
- **Intuition:** Recruiters posting keywords with higher steminess more likely to look for STEM graduates because the activities of the advertised job require STEM knowledge and skills
- Use Bayes' Theorem to link the steminess of keywords in posting *j* to the probability that $j's$ recruiter looks for a STEM graduate:

$$\Pr(STEM|K_j) = \frac{\Pr(STEM, k_1, k_2, \ldots, k_{n_j})}{\Pr(k_1, k_2, \ldots, k_{n_j})}$$

$$= \frac{\Pr(STEM) \cdot \Pr(k_1|STEM) \cdot \Pr(k_2|k_1, STEM) \cdots \Pr(k_{n_j}|STEM, k_1, k_2, \ldots, k_{n_j-1})}{\Pr(k_1, k_2, \ldots, k_{n_j})}$$

$$= \frac{\prod_{k \in K_j} \Pr(STEM|k)}{\Pr(STEM)^{n_j-1}}$$ assuming keywords are posted independently and noting that

$$\Pr(k|STEM) = \frac{\Pr(k) \cdot \Pr(STEM|k)}{\Pr(STEM)}$$

- Estimated as $\widehat{\Pr}(STEM|K_j) = \frac{\prod_{k \in K_j} steminess_k}{\widehat{\Pr}(STEM)^{n_j-1}}$ using smoothed steminess
- Classify j as STEM if $\widehat{\Pr}(STEM|K_j) > \widehat{\Pr}(Non-STEM|K_j)$

Classifying Jobs: evaluating performance

- Out-of-sample experiment design:

250,000 unique random vacancies
from sample with explicit discipline requirements

Training Sample
200,000 vacancies

Test Sample
50,000 vacancies

- Evaluate performance on the **test sample** with a **confusion matrix**:

True / **Predicted**	**Non-STEM discipline required**	**STEM discipline required**
Non-STEM job	Correct classification	Misclassified into Non-STEM
STEM job	Misclassified into STEM	Correct classification

- Evaluates how our classification approach (supervised) performs on unseen data & re-creates the situation where steminess cannot be estimated for all keywords

Classifying Jobs: out-of-sample performance and benchmarking

Replicate experiment 50 times, averages & bootstrapped s.e. in brackets:

	% Correctly classified	% Misclas. into STEM	% Misclas. into non-STEM	Computing Time (hh:mm:ss)	Computer Memory (Giga)	% of Failed experiments
Multinomial Naive Bayes	89.60 [0.138]	9.22 [0.221]	11.62 [0.201]	00:05:44 [00:00:48]	4.54 [0.001]	0
Logistic Regression (Mean & Max steminess)	89.53 [0.134]	9.71 [0.198]	11.26 [0.191]	00:05:35 [00:00:43]	4.70 [0.001]	0
Logistic Regression (~7000 Keywords)	87.16 [0.176]	6.39 [0.332]	19.50 [0.562]	04:57:26 [00:44:20]	14.91 [0.046]	0
Linear Discriminant Analysis	89.95 [0.140]	7.77 [0.212]	12.41 [0.277]	08:31:57 [00:59:47]	95.79 [6.645]	36
Support Vector Machines	90.24 [0.128]	6.59 [0.211]	13.04 [0.237]	09:25:42 [00:51:54]	14.81 [0.705]	2
Tree	72.92 [0.410]	2.65 [6.578]	52.26 [6.725]	04:05:38 [00:36:51]	52.46 [0.490]	8
Boosting Tree	77.04 [1.763]	3.03 [1.047]	43.50 [4.425]	05:43:40 [01:00:04]	56.10 [3.308]	16

Classifying Jobs: Steminess vs. Keywords

Algorithms using keywords directly are:

- **computationally more complex**
  - high dimensionality and sparsity of the "vacancy-keywords" matrix (cf. Manning et al. 2009, Friedman et al. 2008)
  - several methods fail completely: e.g. kNN (nearest neighbours numerous but not "close to the target point")
  - regularization does not help: optimal penalty close to zero, sparsity remains problematic even if remove least frequently posted keywords
  - more efficient implementation?
    RTextTools by Boydstun et al. (2014) employs optimized algorithms from *SparseM* (Koenker and Ng, 2015)

- **less intuitive**:
  - based on dividing the input space into STEM & non-STEM regions with linear (logistic, LDA) and non-linear (SVM) decision boundaries or splitting rules summarized in trees…
  - treat all distinct keywords as completely *separate* dimensions, e.g. "*Budgeting*" as close to "*Java*" as to *"Budget Management"* or "*Costing*"

Using steminess solves these problems:

- **"vacancy-keywords" matrix not needed** – simplifies model & saves computing power
- steminess of *"Budgeting"* (34.41%) much more similar to "*Budget Management*" (36.20%) and to "*Costing*" (52.28%) than to "*Java*" (95.13%)
- **Intuition:** Recruiters posting keywords with higher steminess more likely to look for STEM graduates

Classifying Jobs: Including Job Titles

- 100% of all postings have **job titles**, e.g.: *"Principal Civil Engineer", "Uk And Row Process Diagnostic Business Manager", "Nurse Advisor"...*
- Process the job titles to increase classification accuracy & no. of classifiable vacancies
- Several **Natural Language Processing** steps implemented using R packages *quanteda* (Benoit), *tm* (Feinerer et al.), *stringi* (Gagolewski and Tartanus), *NLP* (Hornik), etc.

1. Tokenization: *"Uk - And - Row - Process - Diagnostic - Business - Manager"*
2. Remove punctuation, stop words…: *"uk - row - process - diagnostic - business - manager"*

- **Final classification of 33m UK vacancy postings** (Jan. 2012 - Jul. 2016) based on:

  - 29,831 keywords (classifiable BGT taxonomy had 9,566)

  - Median vacancy: 7 keywords, 100% of all keywords classified

  - NB algorithm with >90% correct classification rates in-sample & out-of-sample

Outline

1. **Data**
2. **Identifying STEM keywords & jobs**
3. **STEM jobs in the UK**

    Occupational & Spatial distributions

    The wage premium for STEM

    The STEM requirements of "Non-STEM" jobs

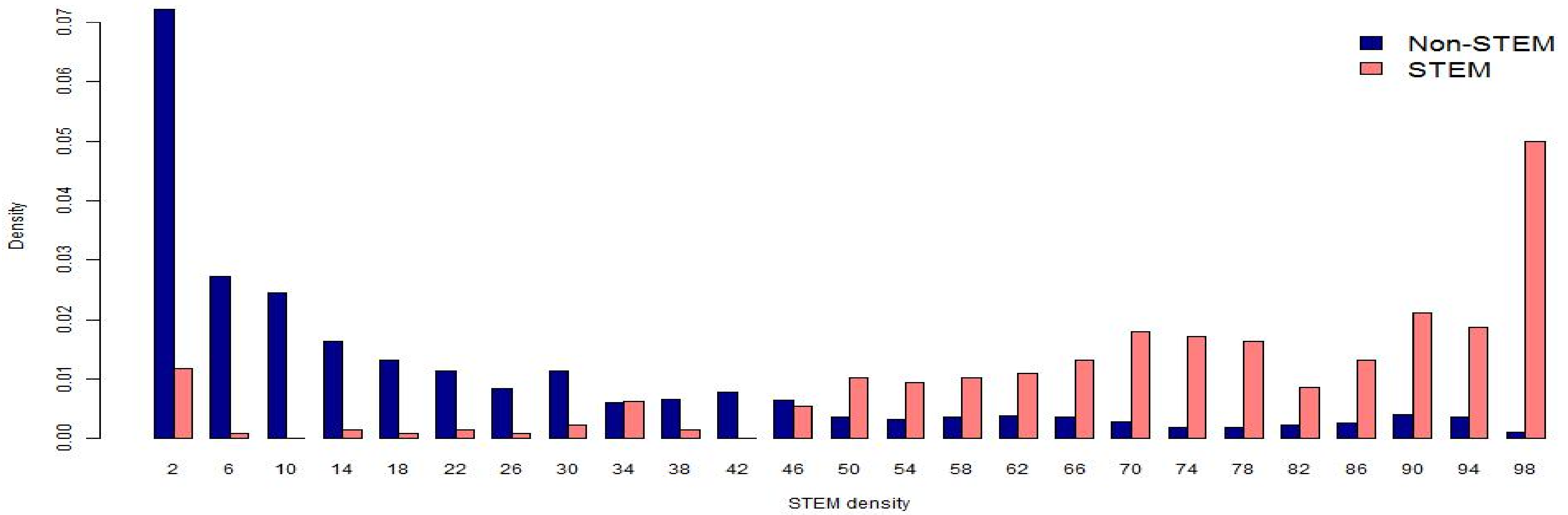
STEM jobs vs. STEM occupations

STEM occupations: - merge lists from UKCES (2015), Mason (2012), BIS (2014) and Greenwood et al. (2011)

- 73 four-digit UK SOC occupations (out of 370, i.e. 20% of all)

	2014	2015	2016 (Jan-Jul)	Total (2012-2016)
No. STEM jobs	1815294	2655532	1865435	10521497
No. STEM jobs in STEM occ.	1172062	1740923	1219474	6885184
No. STEM jobs in Non-STEM occ.	643232	914609	645961	3636313
No. Jobs in STEM occupations	1495158	2146155	1500800	8486364
% of STEM jobs in…				
… STEM occupations	64.57	65.56	65.37	65.44
… Non-STEM occupations	35.43	34.44	34.63	34.56
STEM density of…				
… STEM occupations	78.39	81.12	81.25	81.13
… Non-STEM occupations	13.66	15.27	15.61	14.89

STEM Densities of STEM and Non-STEM occupations



Note: STEM densities in 4-digit UK SOCs. All years combined.

- **Some STEM occupations are not very STEM intense**: *Information technology and telecommunications directors (33.39%), Quality assurance and regulatory professionals (49.94%)* … vs. *Electrical engineers* (99.66%)
- **Diversity of Non-STEM occupations with relatively high STEM densities**: *Business, research and administrative professionals n.e.c.* (46.84%), *Product, clothing and related designers* (45.62%), *Artists* (23.46%)…
- **Finance occupations less STEM intense that often thought:** *Management consultants and business analysts* (25.33%), *Finance and investment analysts and advisers* (7.59%)…

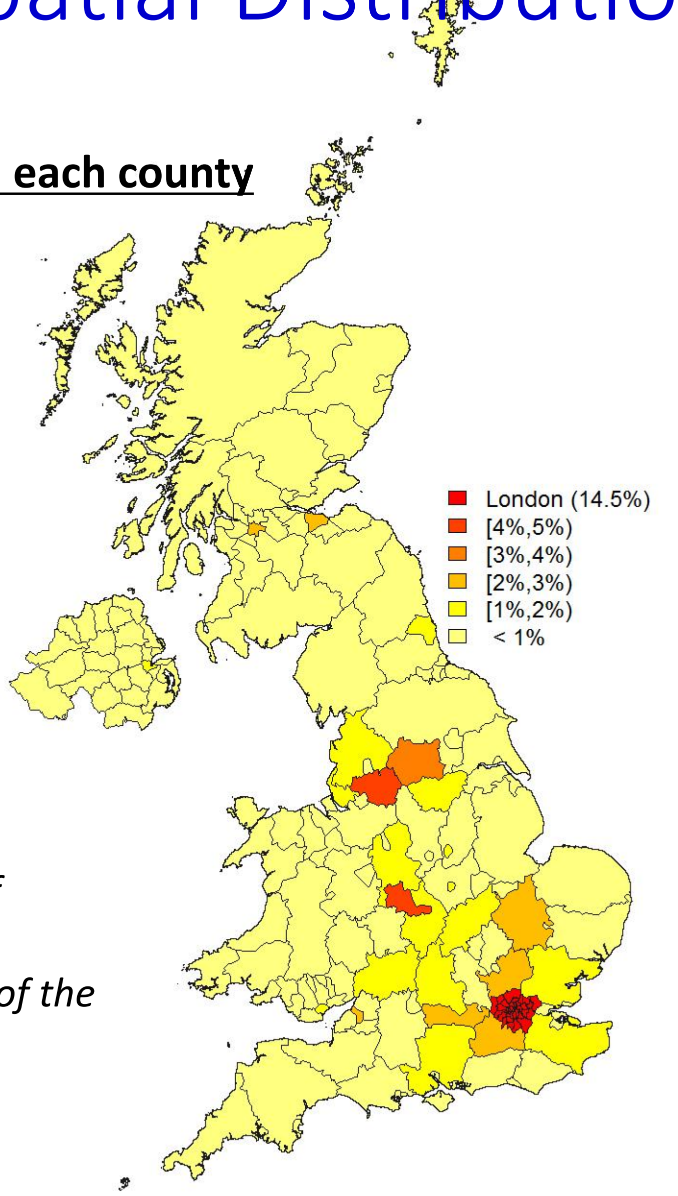
Occupational Distribution of STEM jobs in 2015

"High-level" STEM jobs 74% of all

Major occupational groups	STEM density	% STEM jobs	% jobs in STEM occ.
Managers, Directors and Senior Officials	26.13	7.12	10.41
Professional Occupations	47.9	47.09	51.81
Associate Professional and Technical Occ.	28.59	19.76	23.84
Administrative and Secretarial Occ.	5.47	1.56	0
Skilled Trades Occupations	57.68	12.17	50.04
Caring, Leisure and other Service	3	0.44	0
Sales and Customer Service	10.88	2.32	0
Process, Plant and Machine Operatives	49.49	6.99	0.12
Elementary Occupations	21.45	2.56	0

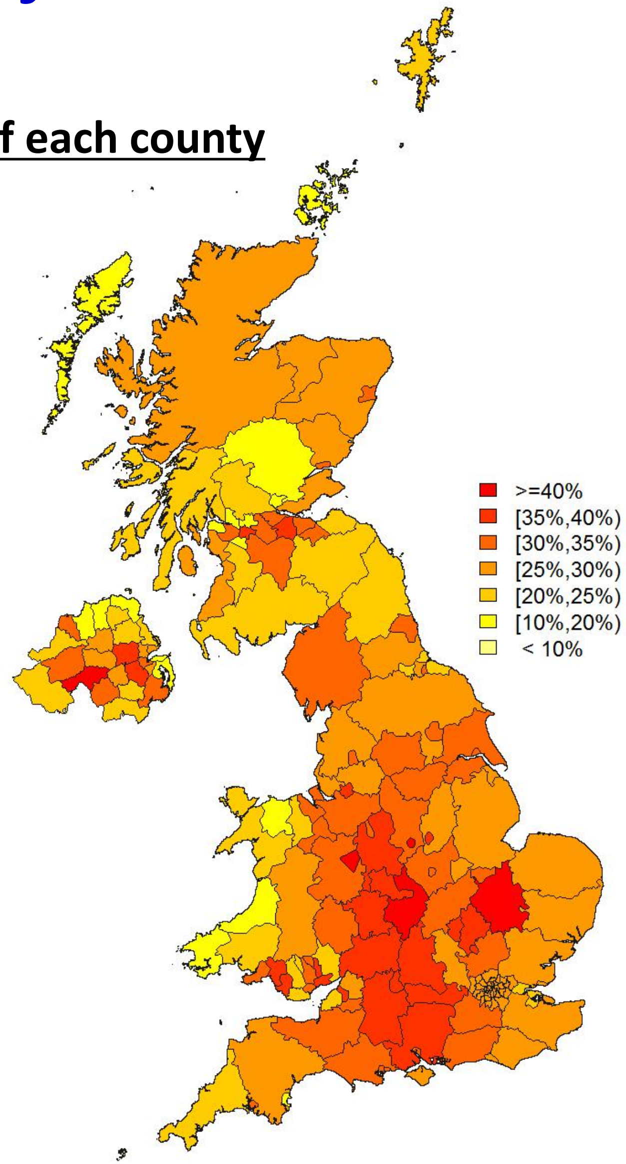
Spatial Distribution of STEM jobs in 2015

<u>% of STEM jobs in each county</u>



Bosworth et al. (2013): London is a *"magnet of STEM workers at the expense of other parts of the country".*

<u>STEM density of each county</u>



Note: Based on the sample of vacancies with County identifiers (77.8% of all vacancies posted). Left map reweighted using ONS ASHE.

The wage premium for STEM

	Dependent variable: $\ln(\textit{hourly salary})$					
	(1)	(2)	(3)	(4)	(5)	(6)
$\widehat{\Pr}(STEM \mid K_j)$	0.319***		0.219***	0.236***		0.125***
STEM occupation		0.293***			0.167***	
$\widehat{\Pr}(STEM \mid K_j)$ **STEM occ.*			-0.047			-0.037
Education						0.049***
Experience						0.030***
London			0.220***			
No. Keywords			0.004***			0.001***
4-digit Occupations	No	No	Yes	No	No	Yes
1/2-digit Industries	No	No	No	No	No	Yes
Counties	No	No	No	No	No	Yes
Year & Month *Pay frequence & Salary Type*	No	No	Yes	No	No	Yes
Clustered s.e.	No	No	Yes	No	No	Yes
Observations	19,856,575			222,451		
Adjusted R^2	0.059	0.053	0.443	0.038	0.020	0.497

*p<0.1; **p<0.05; ***p<0.01

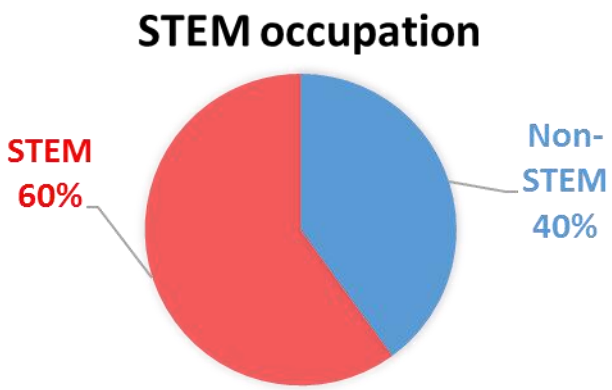
The STEM requirements of “Non-STEM” jobs

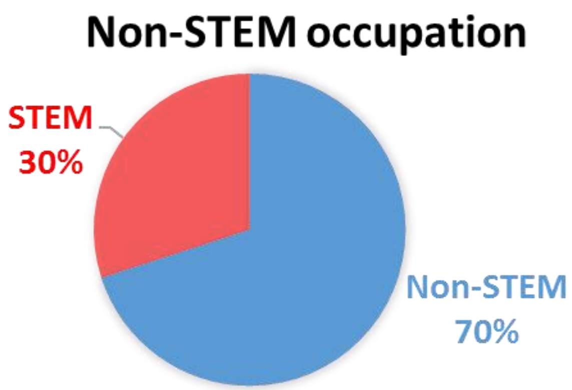
- The STEM knowledge & skills required for the “non-STEM’’ STEM jobs go beyond *‘Problem Solving*’ and ‘*Analytical Skills*’, but very often can be aquired with less than a full time STEM degree:

  “C++”, “3D Modelling”, “Digital Design”, “Big Data”, “Web Site Development”, “jQuery”,…

- STEM recruiters in Non-STEM occupations wish to combine STEM with non-STEM to a larger extent than STEM recruiters in STEM occupations - **hybrid jobs**

<u>% of STEM vs. Non-STEM keywords in a STEM posting (medians)</u>





Conclusion

Contributions

- Debate in the UK: "STEM pipeline leakage'' = wastage of resources?
- New approach to identifying STEM jobs through the keywords posted in online job ads
- Analysis of STEM jobs in the UK: occupational & spatial distributions, wage premium for STEM, STEM requirements of "Non-STEM" jobs

Findings & policy implications:

- "STEM pipeline leakage'' less problematic than typically thought because a significant proportion of recruiters in "Non-STEM'' occupations require & value STEM knowledge & skillls
- However, may still be problematic because:

  - nothing prevents STEM graduates to take up non-STEM jobs within non-STEM occupations

  - a more efficient way of satisfying STEM demand in non-STEM occupations would be to teach more STEM modules in non-STEM disciplines since many of the STEM requirements of "Non-STEM'' jobs do not require a full-time STEM degree

Appendix

STEM jobs vs. STEM occupations
Sample with explicit discipline requirements

Table 2: STEM jobs in the sample with explicit discipline requirements

STEM job =	*% STEM disciplines >50*		*% STEM disciplines = 100*	
	% of jobs that are STEM	% of STEM jobs belonging to	% of jobs that are STEM	% of STEM jobs belonging to
STEM occupations	81.64	69.46	78.45	70.63
Non-STEM occupations	24.11	30.54	21.92	29.37

Notes: Based on the sample of 3957387 vacancies with explicit discipline requirements and an occupation identifier. 1869128 STEM jobs, 1590254 jobs in STEM occupations.