Session 1: Measurement
10:45 - 11:25

Chair: Mirko Draca (Warwick-CAGE)

Anna Valero (LSE)
Economic Growth Goes 'Fractal':
The Changing Structure of the UK's High-Growth Economy

Juan Mateos-Garcia (Nesta and ESCoE)
Building and applying a bottom-up industrial taxonomy of the UK economy

Session 1: Measurement (Continued)
12:30 - 12:50

Daniel Rock (Wharton) (Virtual)
Work2Vec: Learning a Latent Representation of Labor Demand

Centre for Economic Performance

Economic Growth Goes 'Fractal': The Changing Structure of the UK's High-Growth Economy

Anna Valero, CEP and POID, LSE

(with Mirko Draca, Max Nathan, Viet Nguyen, Juliana Oliveira-Cunha, Anna Rosso and Shengxing Zhang)

Modelling an Evolving Economy, 7th October 2022

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

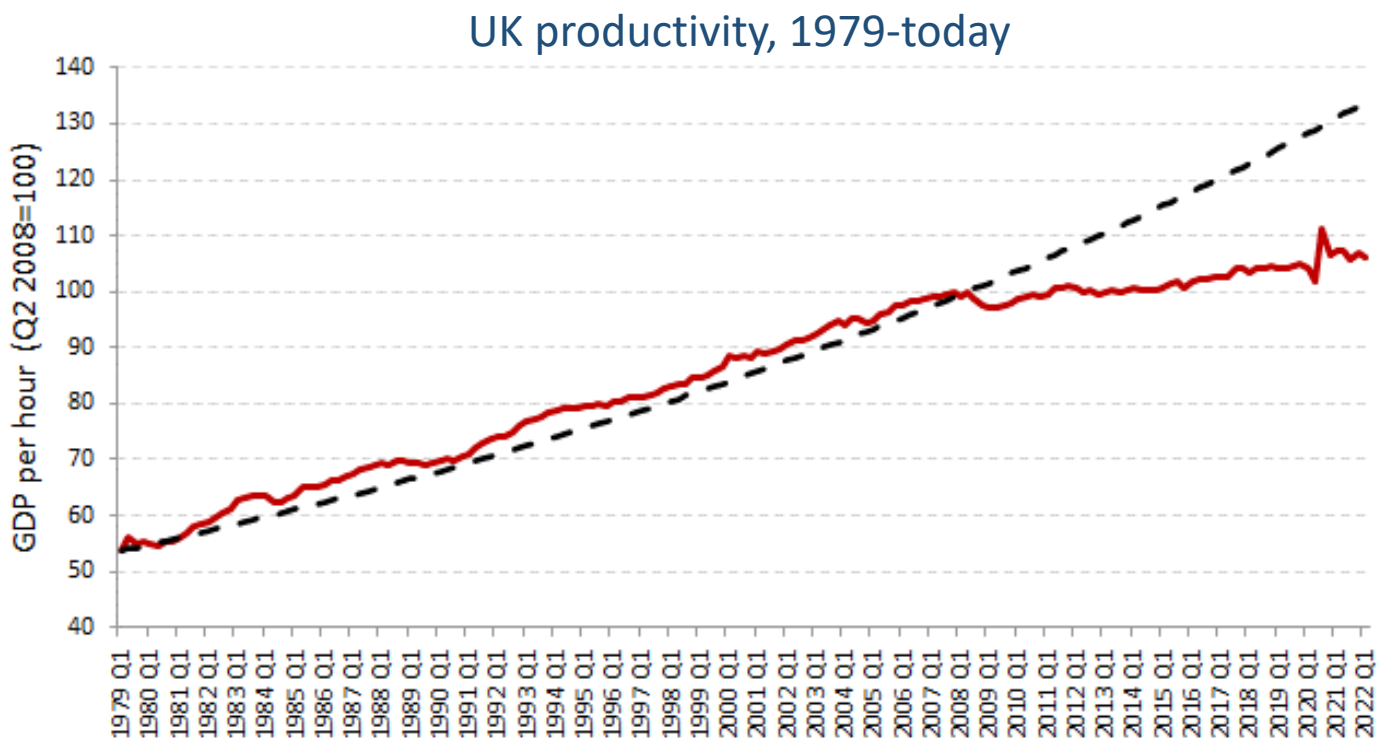
poid Programme on Innovation and Diffusion

UKRI Economic and Social Research Council

Overview

- We study the structure of the UK's 'high-growth' economy using comprehensive data on firm activities and performance from Beauhurst
- We create measures of textual similarity between firms which enable us to locate them in clusters and networks
- We then relate such features to firm outcomes
- **Key findings:**
  - We document a 'fractal' structure amongst high-growth firms – which are split into meaningful clusters
  - High-growth firms appear to be getting more differentiated from each other
  - Originality pays – up to a point – there are better outcomes for firms doing something new, but doing so amongst peers

Motivation 1: High-growth firms are an important source of much needed growth in the UK



Source: ONS Output per hour worked, release date 7 July 2022. Table 32.

Policy paper

Patient Capital Review

The Treasury has now concluded its Patient Capital Review, which considered how to support innovative firms to access the finance that they need to scale up.

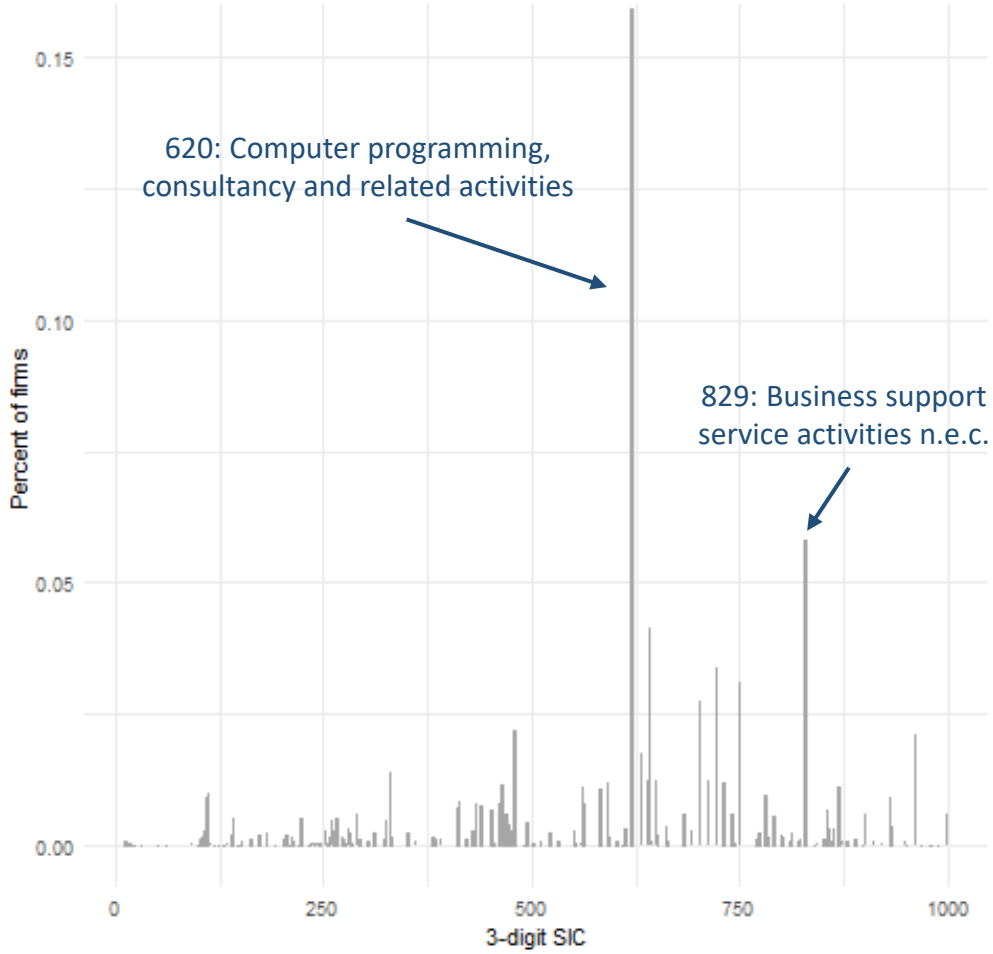
From: HM Treasury and Department for Business, Energy & Industrial Strategy
Published 23 January 2017

Motivation 2: Traditional datasets are limited in their ability to shed light on high-growth firms

- Typically we see revenue or employment growth, but not much about growth potential or intention
- And SIC codes are not very informative on emerging technologies

"… the changing structure of the economy means that SIC will constantly lag reality, under-representing newer industries and over-representing ones that are declining in importance."

Distribution of Beauhurst firms across SIC



Related literature

- **High-growth firms and business dynamism:**
  - Haltiwanger et al. (2017), Decker et al. (2016), Calvino et al. (2018), De Loecker et al. (2022) [Deaton]; Oliveira-Cunha et al. (2021) [Economy 2030]
- **Entrepreneurship and venture capital:**
  - Guzman & Stern (2020), Ewens et al. (2018), Dalle et al. (2017), Kerr et al. (2014)
- **Classifying firms based on textual information**:
  - Competitive dynamics in product markets, firm similarity: Hoberg & Phillips (2010, 2014, 2016), Menon et al. (2018) [public firms, 10-K forms]
  - Start up strategies: Guzman & Li (2022) [start-up websites vs incumbent 10-K forms]
  - Mapping emerging sectors using website text: Nathan & Rosso (2015) [digital economy], Mateos-Garcia et al. (2014, 2018) [video games/VR/AR], Bishop et al. (2022) [within SIC]
  - Technological innovation: Kelly et al. (2021) [patents], Kogan et al. (2022) [patents-occupations]

Data

- Beauhurst tracks UK firms that have hit any of 8 triggers since 2011
- Tracking ends upon exit (successful or unsuccessful)
- Comprehensive and curated profile including company descriptions, financials, fundraising activity, outcomes…
- We use the web-scraped company description

Beauhurst

- ✓ Received equity investment
- ✓ Underwent an MBO/MBI
- ✓ Received venture debt
- ✓ Reached scaleup status
- ✓ Received a large innovation grant
- ✓ Spun out from an academic institution
- ✓ Graduated from a selected accelerator
- ✓ Featured on a selected high-growth list

	Web-scraped description	Analysts' description
Deliveroo	Deliveroo is on a mission to transform the way the world thinks about food delivery. It's not a chicken chow mein and a night on the sofa anymore, it's your favourite local restaurant, it's a dinner party, a date. We're five years in, and along the way our team have taken hundreds of ideas from brainstorms to global roll-outs, like Deliveroo Editions & bespoke kitchens designed to host a locally-curated selection of restaurants. Editions are our solution to ensuring that our customers have access to the best of the food-scene, no matter where they live. And that's just what we're like at Deliveroo, no compromise allowed and lots of food-inspired challenges to get your teeth into. Out-of-the-box thinking is actively encouraged and we move quickly to make great ideas happen. We're energetic, fast-paced and blow off steam with free-for-all Friday lunches. It's a formula that's working too & we're bringing great food to customers in 13 countries and over 200 cities.	Deliveroo provides delivery services for restaurants, using technology to predict the time taken to prepare meals and efficient ways of delivering orders using the location of restaurants, customers and riders.

Cleaning up the text

- We apply standard pre-processing steps to construct a vocabulary of unique words

	Mean	Median	Max	Min	Sample total
(A) Beauhurst web-scraped (N = 33,973)					
Total words	76.38	65	1,655	13	
Unique words	62.95	56	826	13	94,298
(B) Beauhurst analysts (N = 33,973)					
Total words	11.58	11	49	1	
Unique words	11.34	11	44	1	41,385

Notes: This table provides information on the amount of text available in different UK data sources. The sample frame is the set of 33,973 BH high-growth firms, obtained after the pre-processing steps

Notes: The top 100 unique words in our BH web-scraped data, obtained after the pre-processing steps. The size of the word reflects the number of times it has been used across firm descriptions.

Methodology overview

Cosine similarities across firm pairs (H&P, 2016):

- Start with vocabulary of words
- Each firm represented by a vector where each element is populated with a 1 if the firm uses a word, 0 if not
- These are converted into frequencies and then stacked into a matrix
- Cosine similarity is calculated between each two firms
- [0,1], higher when two firms use more of the same words
- Similarities stacked into NxN firm matrix

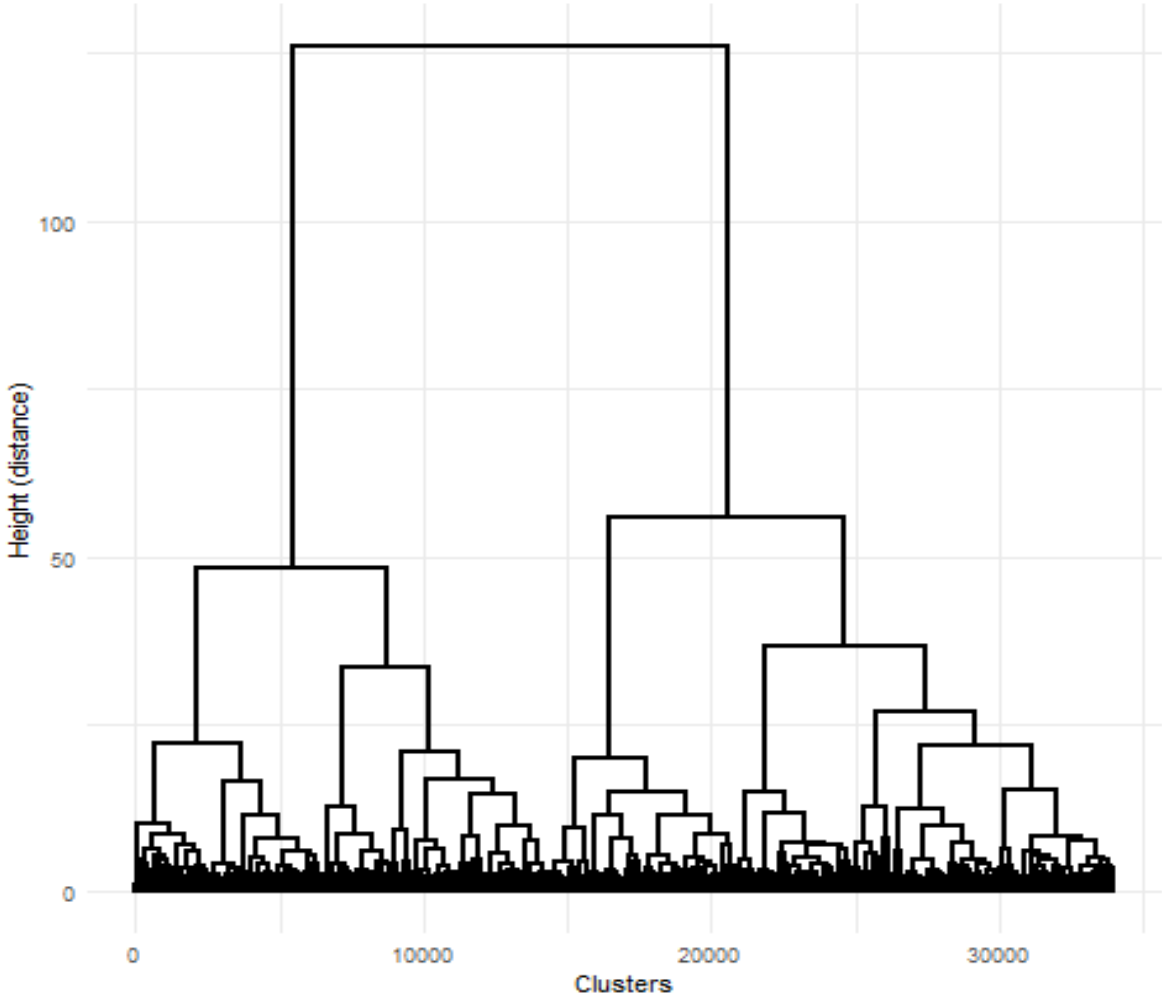
…form the basis of 3 analyses:

- Hierarchical clustering
- Network connections
- Differentiation across and within cohorts

Clustering

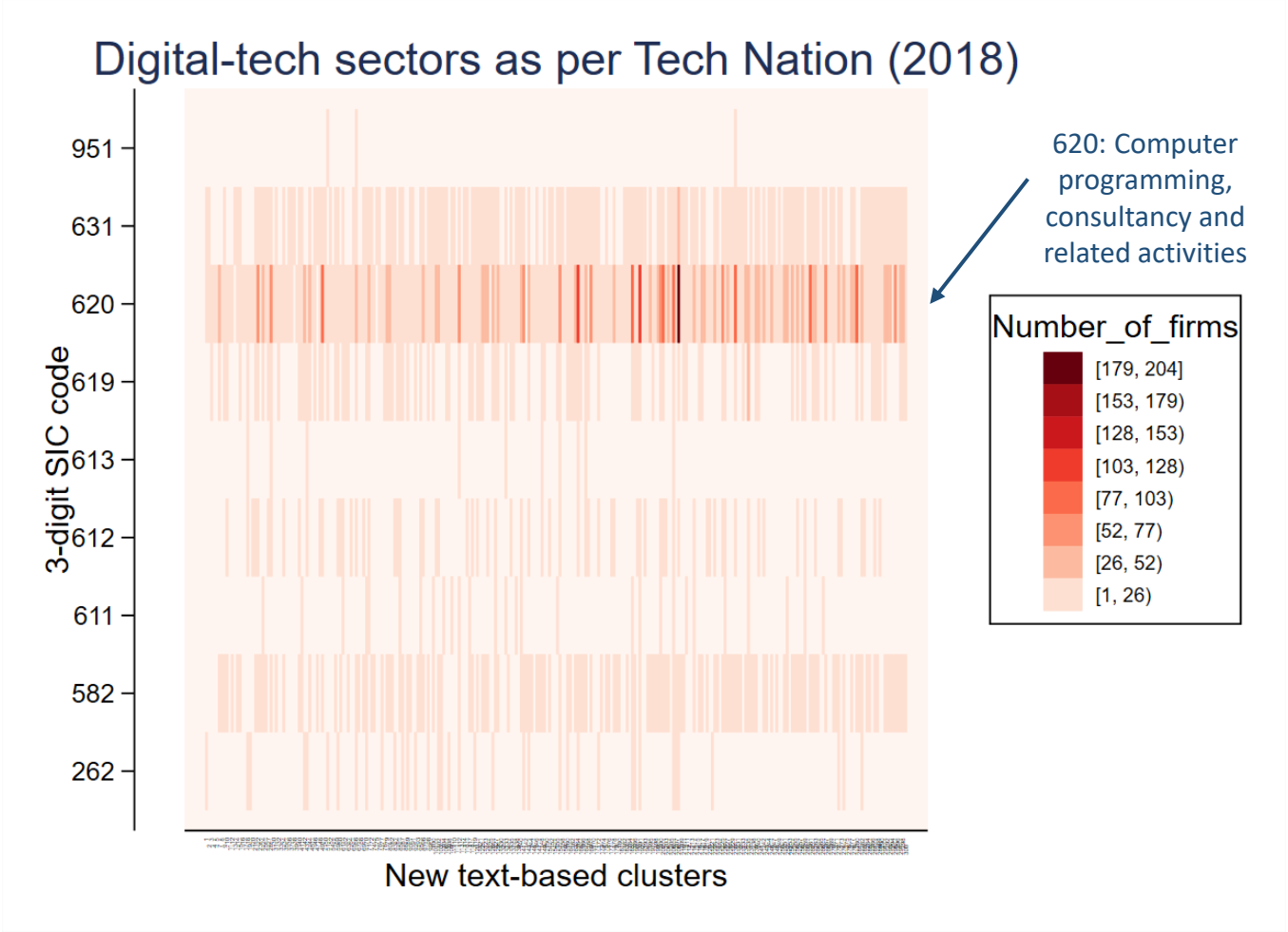
Hierarchical clustering of high-growth firms

- We run a clustering algorithm on the firm matrix
- This creates a branching structure (‘fractal’ groupings of firms)
- We place similar firms into 300 discrete bins (vs 286 3 digit SIC codes in our data)

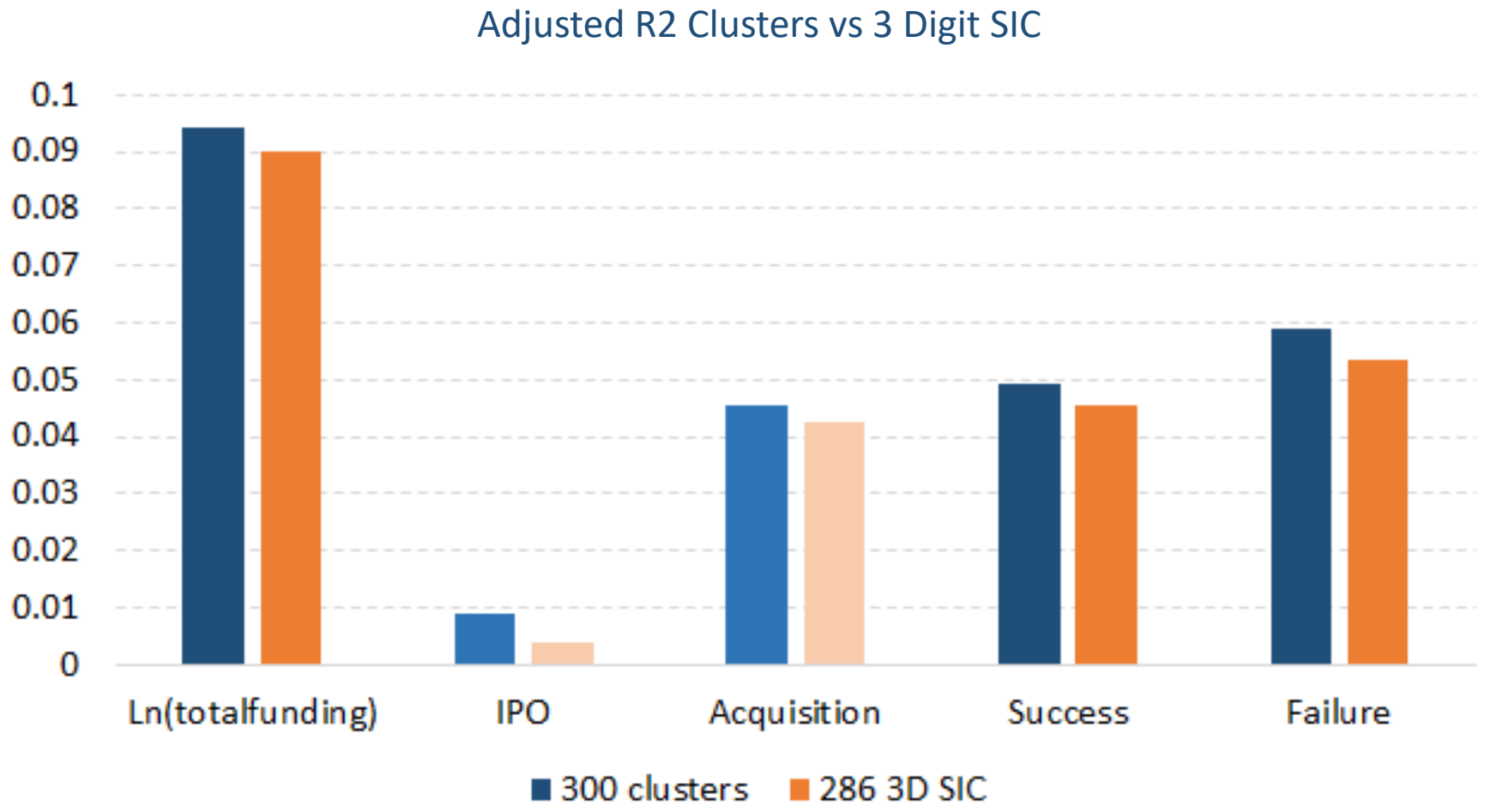


Some of the key sectors are spread across clusters

- Around 20% of sample are in ‘digital sectors’ – mainly 620
- Such codes are split across many clusters (272 out of the 300)



Clusters do a slightly better job of explaining key outcomes

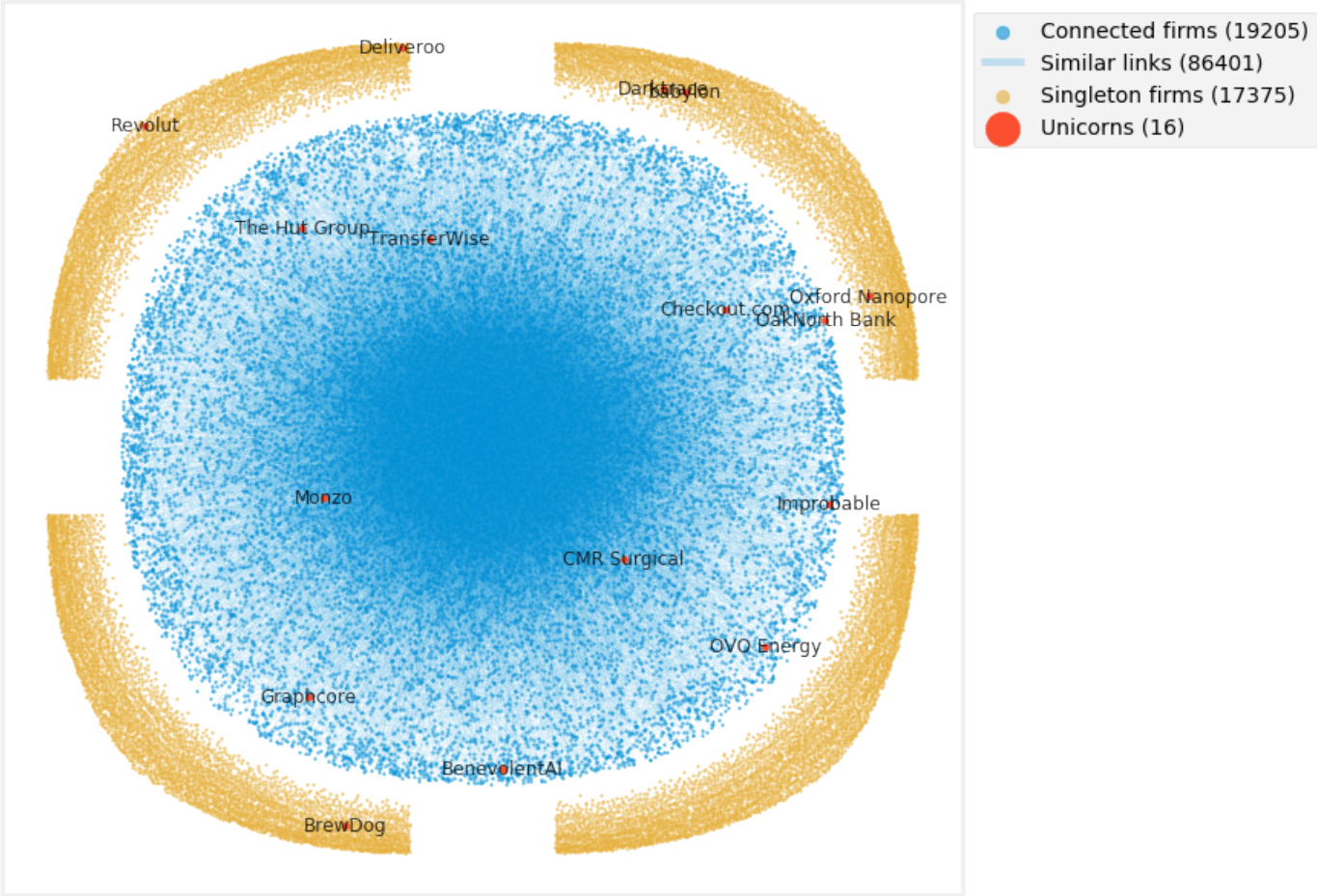


Notes: Regression sample includes Beauhurst tracked companies. Adjusted R2 from OLS regressions of dependent variable on clusters or 3 digit SIC dummies, respectively. Birth cohort dummies included in all regressions. Failure = business death or zombie status

Network connections

Network based structures

- We adopt a threshold of 0.2132 cosine-similarity to define a connection between firms (H&P, 2016)
- "Singleton" firms vs "connected" firms
  - Those closer to the centre are more connected



Notes: Fruchterman-Reingold force directed algorithm to position connected nodes.

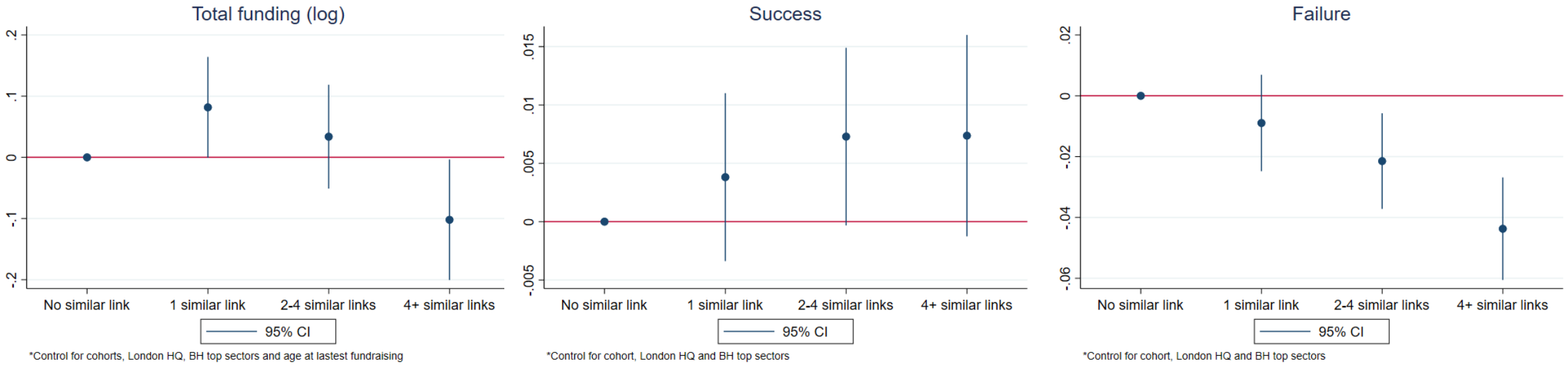
Some connections appear to be a good thing (1)

	(1) Ltotalfund	(2) success	(3) failure
1 similar link	0.0817*	0.00382	-0.00894
	(0.042)	(0.004)	(0.008)
2-4 similar links	0.0338	0.00729*	-0.0215***
	(0.043)	(0.004)	(0.008)
4+ similar links	-0.102**	0.00737*	-0.0437***
	(0.050)	(0.004)	(0.009)
HQ region is London	0.538***	0.00835***	-0.0351***
	(0.031)	(0.003)	(0.006)
Age at lastest funding	0.437***		
	(0.009)		
Cohort	Yes	Yes	Yes
Top sectors	Yes	Yes	Yes
N	12,663	18,231	18,231
r2	0.243	0.0195	0.0375
ymean	12.94	0.0327	0.192

Notes: Sample includes Beauhurst startups founded from 2010-2019. Robust standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

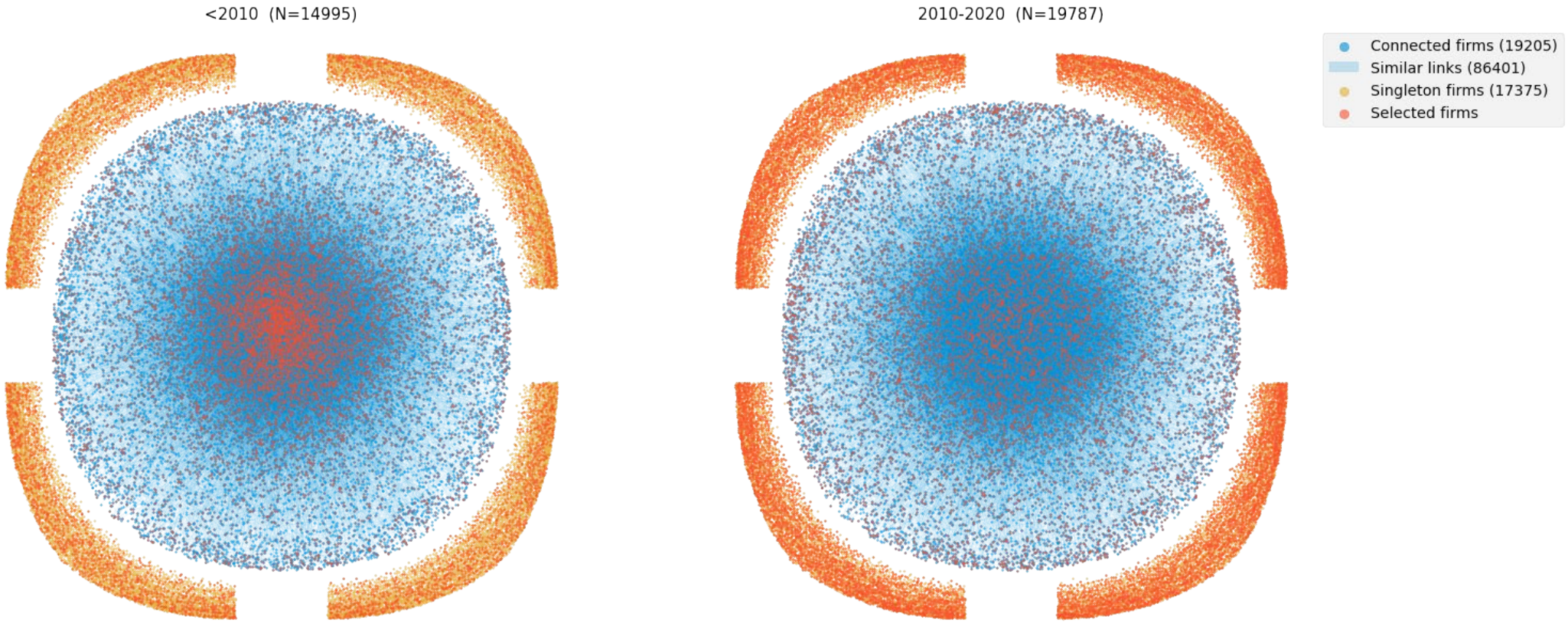
Some connections appear to be a good thing (2)

High growth firm outcomes and network connections



Notes: Sample includes Beauhurst startups founded from 2010-2019.

Across cohorts, firms appear to be less connected



Differentiation across and within cohorts

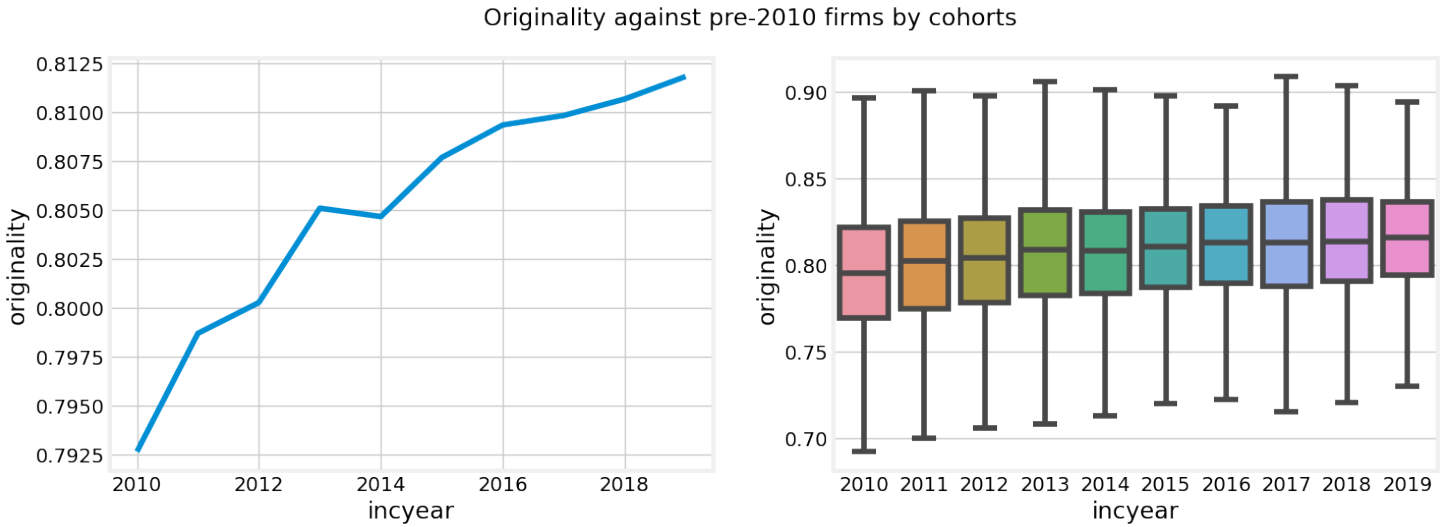
Originality and trendiness measures

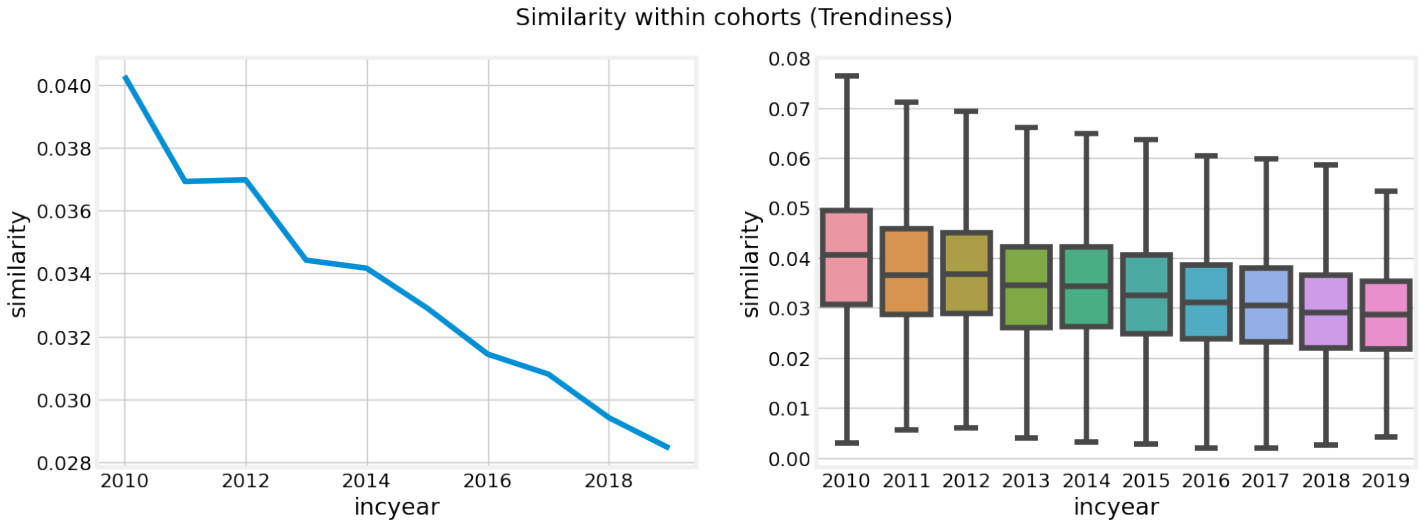
- **Originality**:

1-similarity with the most similar of high-growth start-ups in pre-2010 cohort

- **Trendiness**:

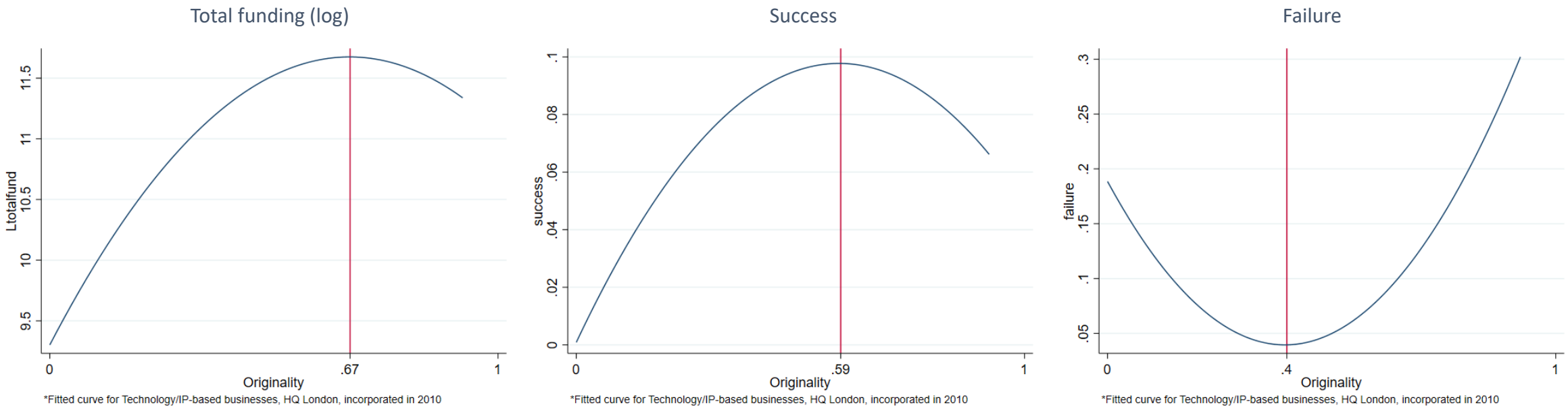
Average similarity with all high-growth start ups in the same birth cohort





It’s good to be different to the past, to some extent

High growth firm outcomes and originality



Notes: Functions implied by coefficients estimated in regressions of specified dependent variable on originality, originality^2, London dummy, cohort fixed effects (and age at last funding). Coefficients on originality and originality^2 statistically significant. Sample includes Beauhurst startups founded from 2010-2019. Robust standard errors in parentheses.

But also it's good to be in a currently 'trendy' area

	(1) Ltotalfund	(2) success	(3) failure
trendiness	5.791***	0.582***	-1.537***
	(1.39)	(0.12)	(0.27)
HQ region is London	0.548***	0.00863***	-0.0348***
	(0.03)	(0.00)	(0.01)
Age at latest funding	0.437***		
	(0.01)		
Cohort	Yes	Yes	Yes
Top sectors	Yes	Yes	Yes
N	12,663	18,231	18,231
r2	0.24	0.02	0.04

Notes: Sample includes Beauhurst startups founded from 2010-2019. Robust standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01

Conclusions and future work

- Text based measures of similarity can form the basis of a number of informative measures about high-growth firms
  - Including new measures of originality ~ innovation
- We find that our clusters are informative
- And high-growth firms appear to be getting more differentiated
  - Originality pays – up to a point – there are better outcomes for firms doing something new, but doing so amongst peers
- Future work
  - Other measures of similarity / networks – looking deeper within SIC codes
  - Other outcomes (e.g. innovation explicitly, stages of fundraising – seed/growth)
  - Dynamics - changes in firm descriptions over time
  - Networks of individuals in the high-growth economy

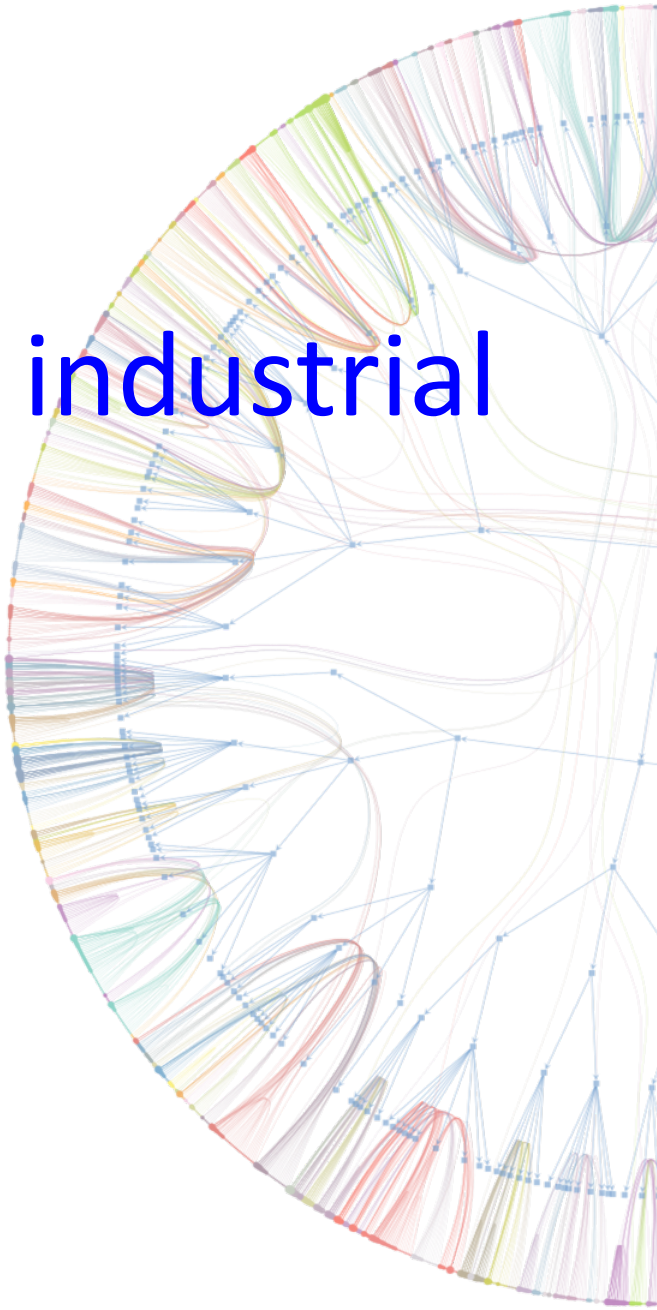
Thank you!

Building and applying a bottom-up industrial taxonomy of the UK economy

Alex Bishop, Juan Mateos-Garcia and George Richardson

ESCoE workshop

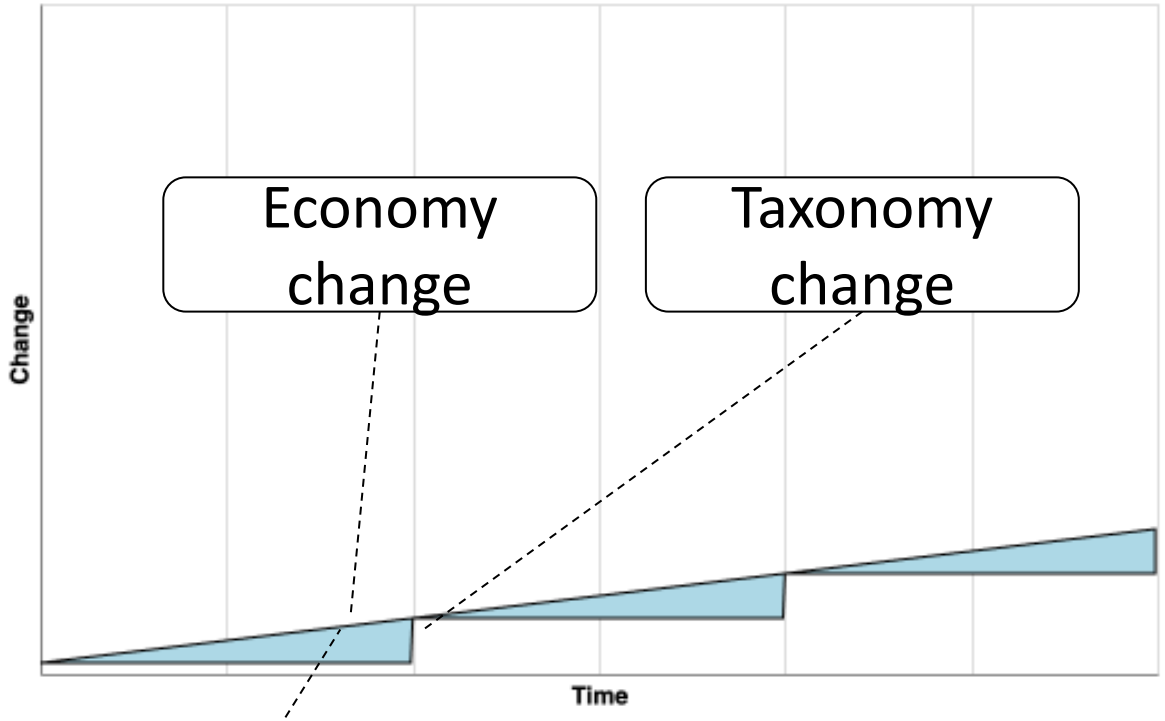
7 October 2022

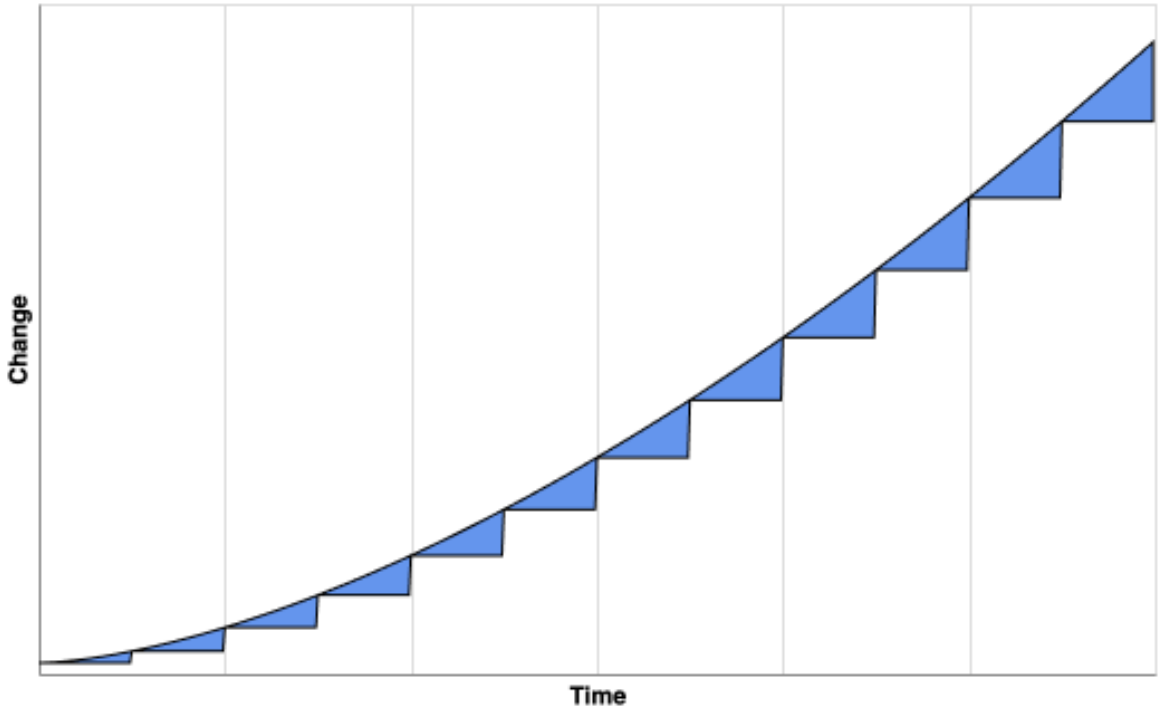


nesta

ECONOMIC STATISTICS CENTRE OF EXCELLENCE

The problem





Measurement gap

Can we use new sources of data and analytics to close this growing gap?

nesta

Our data

1.8 million business websites identified via IP register

Matched with Companies House

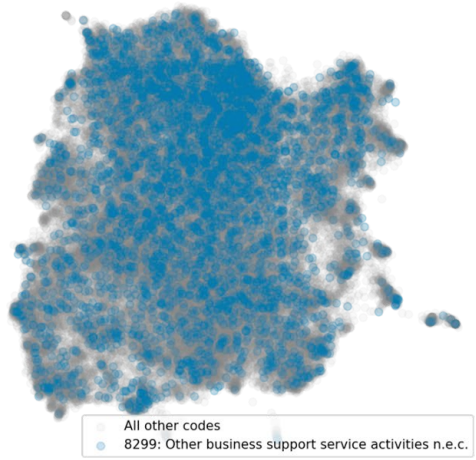
Some sectoral biases: agriculture underrepresented, knowledge intensive and digital overrepresented.

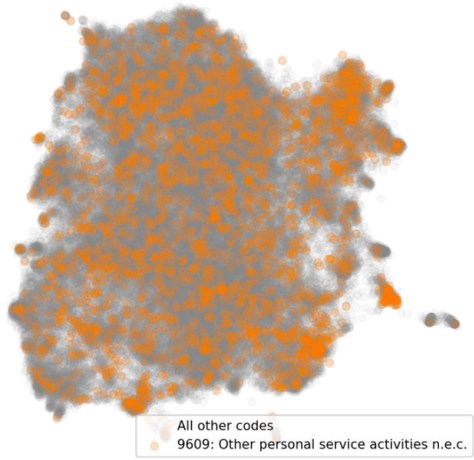
SIC4	Description
6201	We explore successful digital...

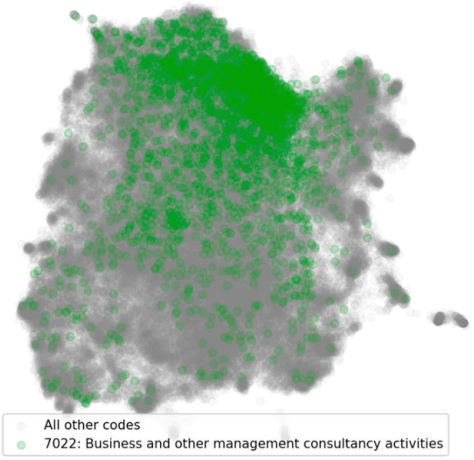
A test

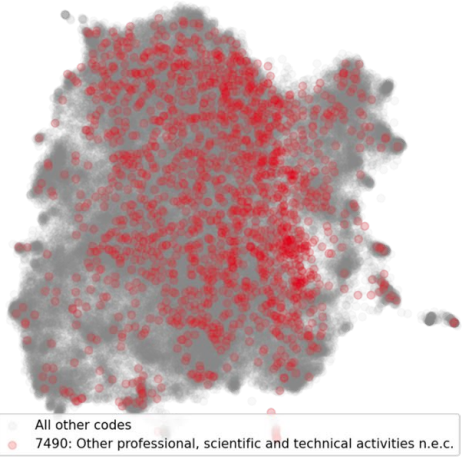
Can we predict SIC codes using company descriptions and state-of-the-art language models?

- Not very well
- Our model struggles especially with SIC codes capturing other not elsewhere classified knowledge intensive activities



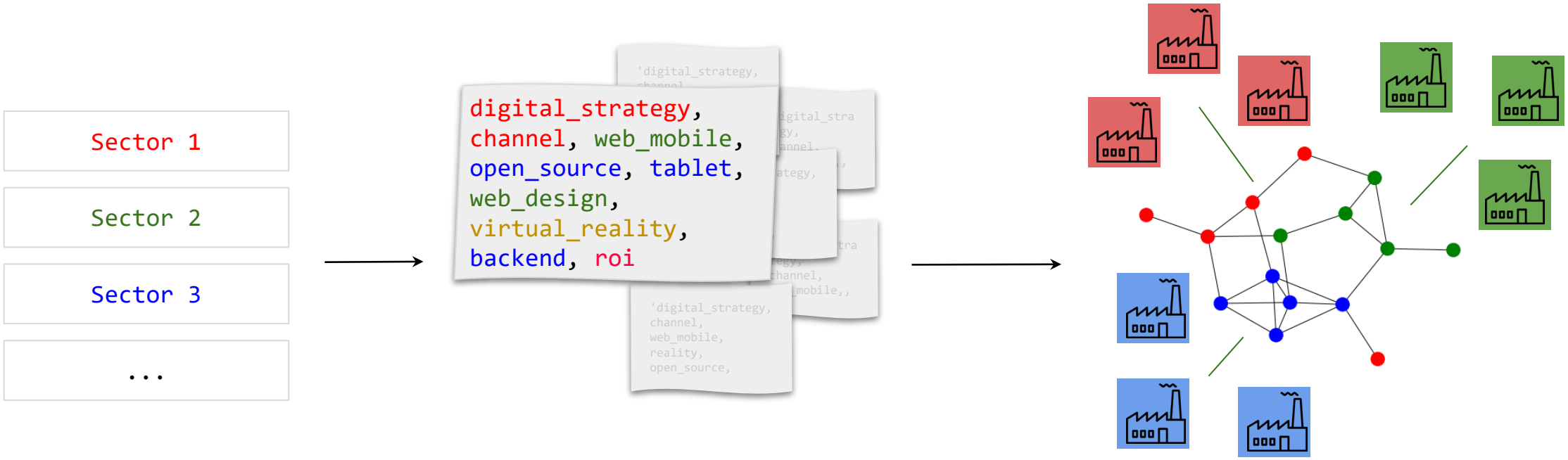






Beyond (or below) SIC codes

Can we analyse the text in company descriptions to generate an alternative, bottom-up taxonomy ("text sectors")?



We perform the clustering *inside 4-digit SIC codes* and then reassign companies to their closest *text sector in semantic space.*

A bonus we haven't used yet is that companies are tagged with multiple text sectors

What comes out

...

6820_21: growth global investor entrepreneur enterprise economy retail strategic global_network real_estate b2b innovation startup invest fund_manager expand capital corporate fund ambition

6910_5: immigration immigration_law visa case law lawyer legal appeal migration law_firm regulate_org represent legal_advice advise human_right settlement assistance complex remain adviser

7111_27: architecture architectural sustainable bim architect idea brief designer aspiration low_energy stage energy architects architectural_practice architectural_design innovation communication int

6820_10: service_user young_people autism volunteer carer begin mental_health dementia disabled_people social woman forum encounter neighbourhood therapeutic therapist person_centre facilitate people_

7410_14: construction_project consultancy_service engineering_consultancy project_management architectural company_formation sub_contractor feasibility new_build co_ordinate consulting procurement nut

8690_22: tooth acupuncture pain treatment injury condition therapy sport_injury body exercise patient nail skin movement clinic rehabilitation mechanism non_invasive treatment_plan muscle

8690_0: volunteer mental_health woman funding tennis patient spa fund carer begin counselling prison branch premise open_date expand room open therapy pool

8299_97: clock item purchase store stock buy golf ship shop packaging print budget retailer baby machine antique merchandise coin gun quality_product

6920_14: accountant accountancy accounting firm_accountant accountancy_service tax tax_return accountancy_taxation sole_trader member_org accountancy_firm taxation tax_advice client_alike company_form

6202_61: club accessibility enclosure membership class tourist visitor room self_cater garden food estate café restaurant pink island garden_room perform luxury list

7499_7: bird horse animal wildlife barn balloon stable countryside farm specie breed timber conservation rug travel dog population ambassador trip habitat

4322_32: electrical air_conditioning underfloor_heating test design_installation install gas pump roof design_install pipework pipe instal injection heat_pump mechanical refrigeration conditioning con

6820_37: beach cottage accommodation bed_breakfast harbour island hotel self_cater coast explore stone_throw famous bedroom village time_drive room away short_walk view situate

4110_28: conservatory roof window kitchen garden door space glass extension room exterior instal finish transform_home outdoor style window_door option indoor apartment

7022_71: insolvency debt tax account restructure claim cash_flow finance risk recovery disclosure funding performance business_owner run_business face advise audit turnaround increase

5610_28: food food_drink food_beverage bakery vegetable manufacturing restaurant manufacturer kitchen factory operation packaging door innovation mindfulness supply_chain sandwich ingredient growth ma

4321_21: automation home_automation gate robotic manufacturer access_control programming design_install audio_visual meter install electrical installer reliable instal testing automate electronic late

8899_19: trustees company_limit volunteer fund forum grant scheme society drug begin village brain_injury foundation register county found local_community payment educational advocacy

4791_63: kitchen garage make_model manufacture equestrian marque main_dealer dealer warranty mechanic car_servicing mot manufacturer fleet factory equine glazing late_diagnostic specification car_deal

5911_18: film video_production video content animation corporate_video post_production shoot short_film video_content crew digital filmmaker documentary studio editor medium message visual television

...

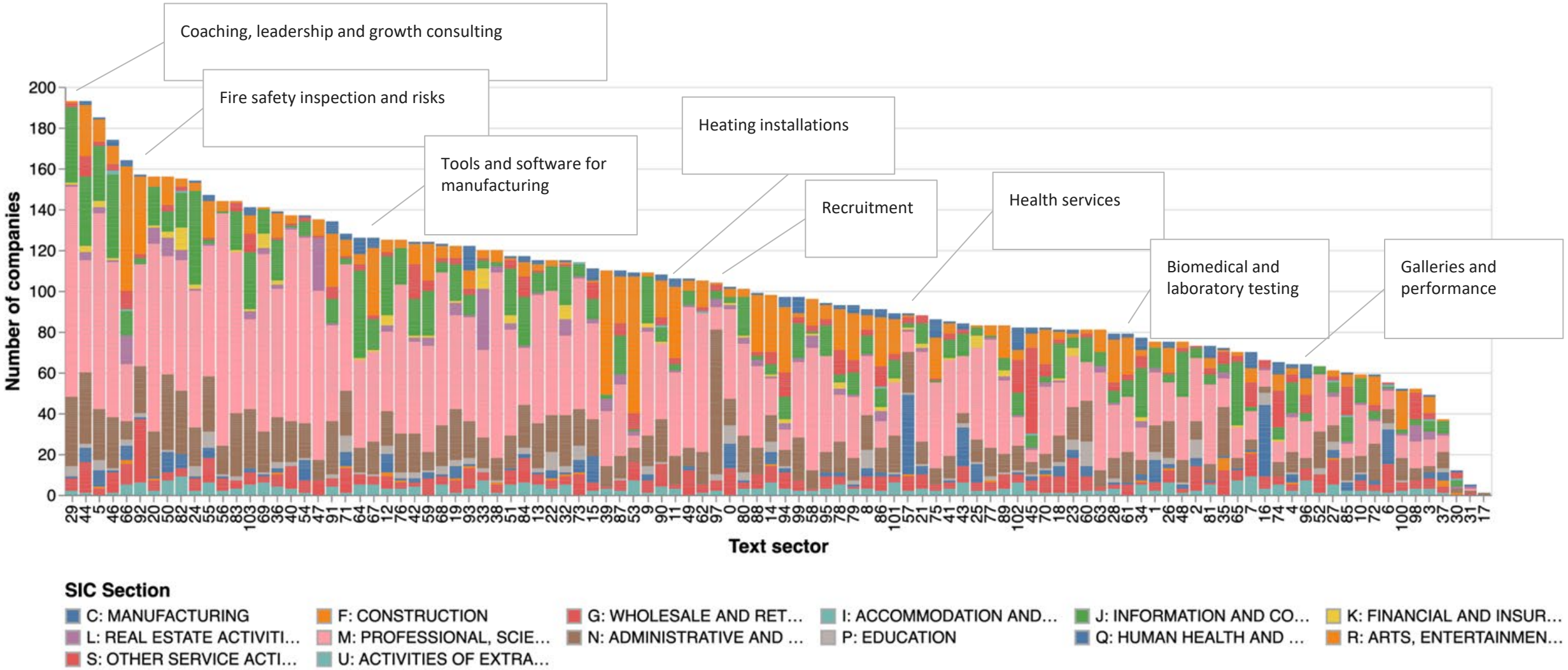
nesta

There are 1800 more text sectors where these came from

Unpicking messy sectors

7490: Other professional, scientific and technical activities n.e.c.

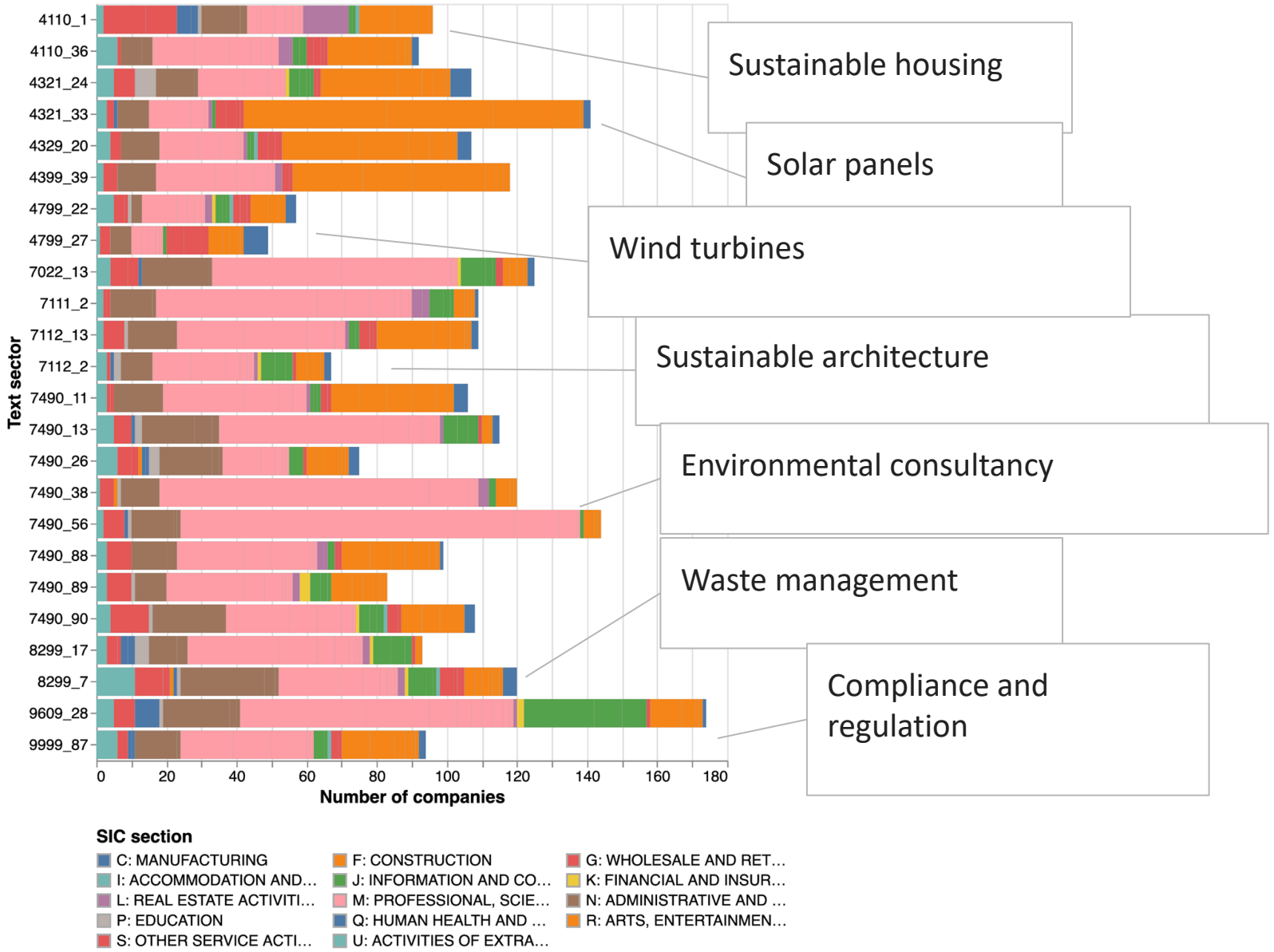
10,746 companies in 104 text sectors. 85% reassigned from other SIC4s.



Defining policy-relevant sectors

We can also use our outputs to study industrial activities in sectors of interest: eg. study ‘the environmental sector’.

- We identify 2,508 firms in 27 text sectors with names that mention 3+ terms related to the environment, sustainability, renewables etc.
- These companies are spread across 42 SIC4 codes.

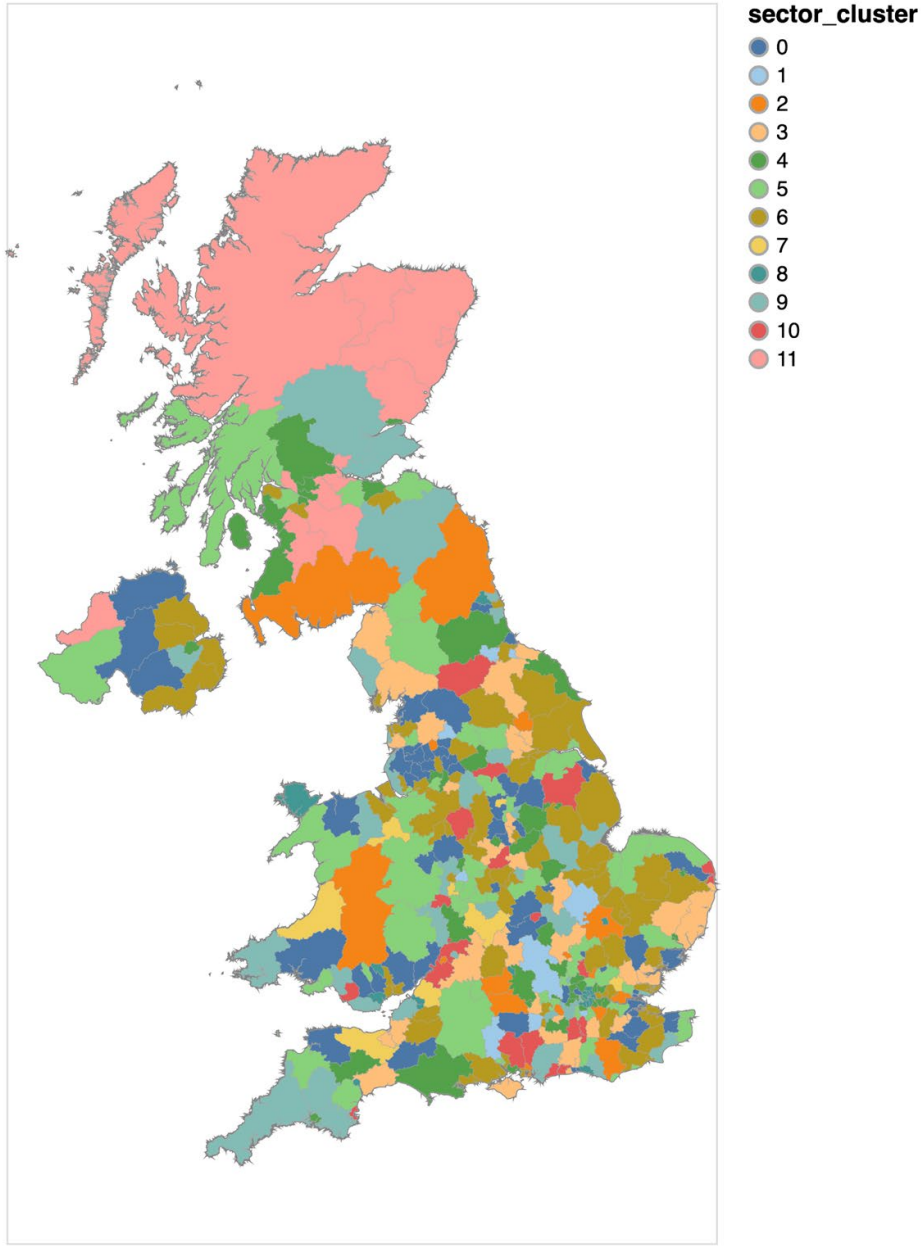


nesta

Analysing economic geography

Can this new taxonomy help us to characterise the economic geography of the UK?

- Measures of economic composition (complexity) based on text sectors are more strongly associated with local economic outcomes than those based on SIC codes.
- In forthcoming work, we cluster local authorities based on their text sector composition and analyse differences between them in secondary data.



nesta

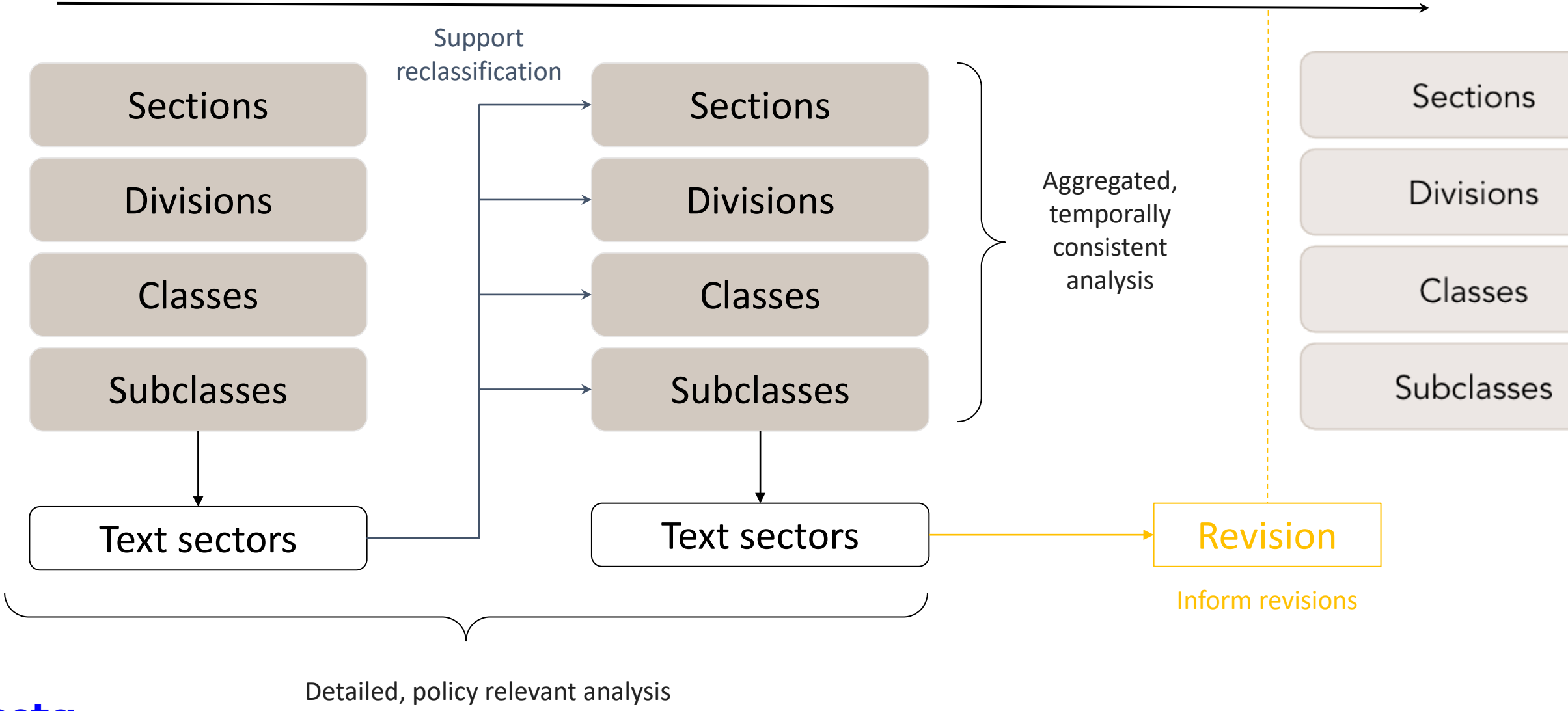
Analysing economic geography [2]

Less productive and knowledge intensive LAs across the country (including the South)

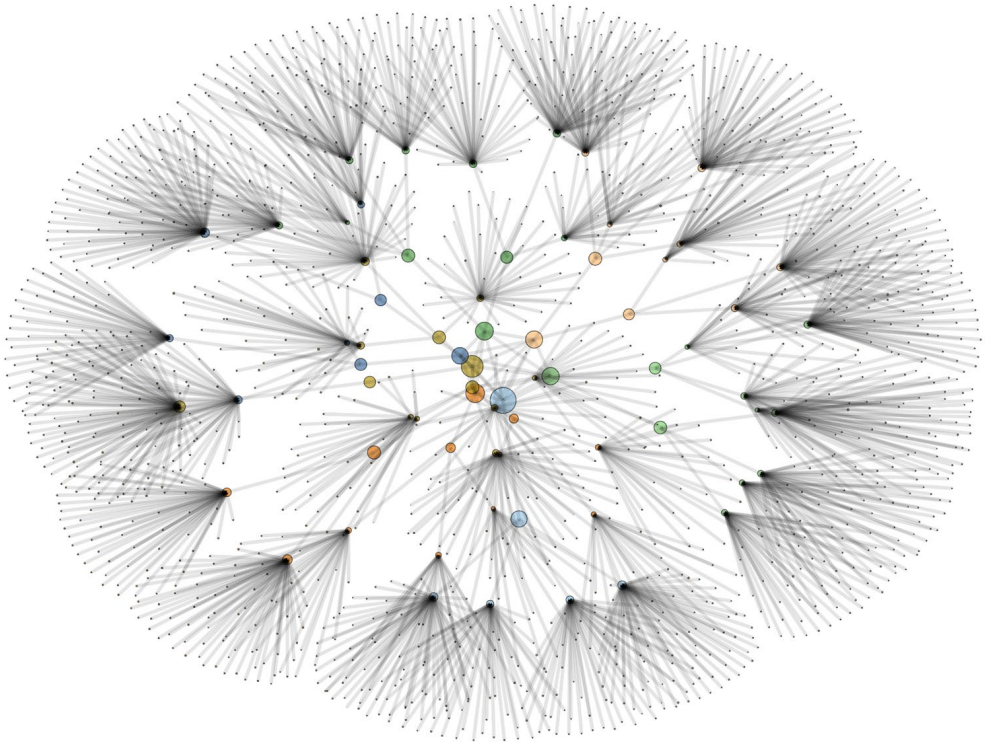
Wealthy and innovative LAs / boroughs in London and the South East

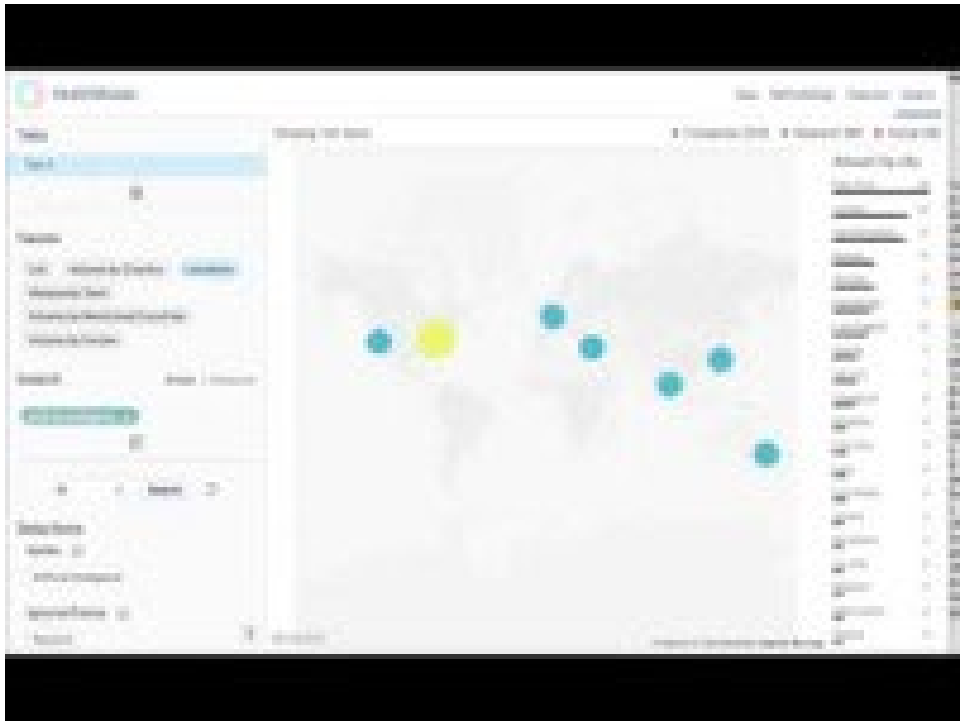
nesta

How could one implement this?



How would you make it useful?





Current version of the taxonomy: striking but opaque

Interactive and explorable

Our bottom-up industrial taxonomy based on web data could be used to complement and augment existing top-down taxonomies

- Process innovation: identify misclassified companies, inform revisions
- Product innovation: inform new policy analyses, help develop new types of policies
- This targeted approach could help bypass mitigate risks created by biased, noisy web data

Need to do this openly and transparently

- https://github.com/nestauk/industrial-taxonomy

Thank you

Juan.mateos-garcia@nesta.org.uk

https://twitter.com/JMateosGarcia

nesta