Hackathon on data science for STI policy

STIP Lab and OECD – TIP event

Research question:

To what extent is it possible to characterise typologies of policy proposals on the theme of scientific employment and research careers?

7 June 2022

SPRU team

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

EQUIS ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Objectives and Agenda

Objectives

To identify key themes related to scientific employment and research careers in available policy proposal databases, so to support the decision-making process

Agenda

1. Methodology and scope of analysis
2. Overview of the target policies
3. Results by themes
   - A. Research career
   - B. Gender balance and inclusiveness
   - C. Inter-sectoral mobility
4. Conclusions

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

Research Methodology

Key points

1. The analysis is based on the EC-OECD STIP-Compass dataset

2. The analysis selects 3 themes under the topic of human capital and careers (HR policy) as our main focuses
   (TH44_Inter-sectoral mobility, TH53_Research careers, TH54_Gender balance and inclusiveness)

3. Descriptive analysis and text mining
   (Principal component analysis (PCA) and k-means clustering)

4. Interpretation and conclusions build on integrating the quantitative analysis and a literature review

Research Methodology

Text Mining (PCA and clustering analysis)

1. Extract the text of HR policies

   (Extract text information is from “ShortDescription” and all “Objectives” in the dataset)

2. Delete words according to frequency (the most and the least used), delete meaningless words manually

   (e.g. one, two, also…)

3. Get tfidf (term frequency–inverse document frequency)

4. Conduct PCA analysis (2 components that explain around 20% information of the whole text.)

5. K-means Clustering

   1. Select k value: Elbow method, Average silhouette method, Gap statistic method, PCA approach and hierarchal clustering
   2. Conduct k-means clustering, compare results by term frequency in each cluster
   3. Connect results with literature review

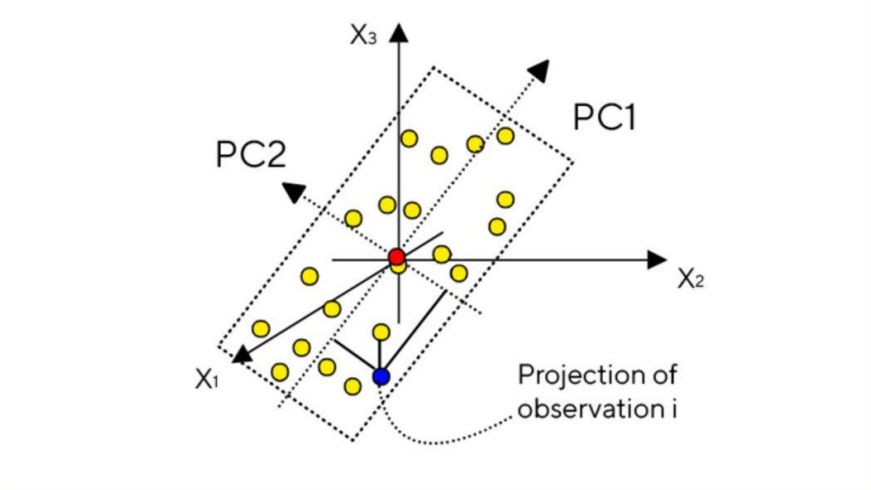
Research Methodology

Text Mining (PCA and clustering analysis)

The retionale of PCA and K-means clustering

1. PCA (Principal component analysis)

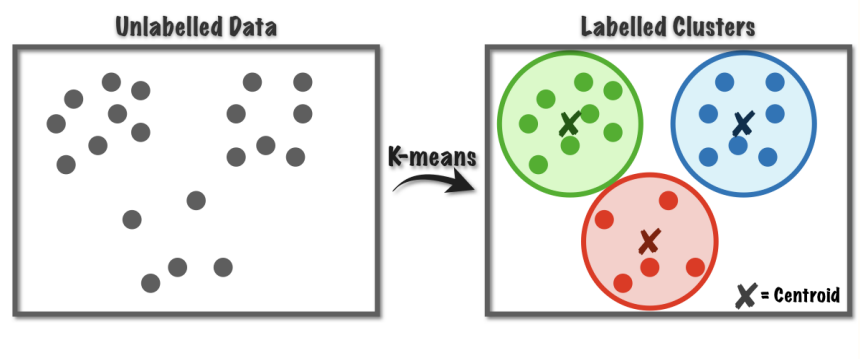
Consider the total variance in the data, and transforms the original variables into a smaller set of linear combinations, and it is easy to find the structure of text data.



PCA Approach

2. K-means Clustering

Aim to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. *k*-means clustering minimizes within-cluster variances.



K-means Approach

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

Research Scope: Human Capital Themes (STIP database)

Human Resources for Research and Innovation policy

6 Themes

- Inter-sectoral mobility (TH44, 113 policies);
- STI human resource strategies (TH50, 240 policies);
- Doctoral and post-doctoral research (TH52, 296 policies);
- Research careers (TH53, 285 policies);
- Gender balance and inclusiveness (TH54, 238 policies);
- International mobility of human resources (TH55, 313 policies);

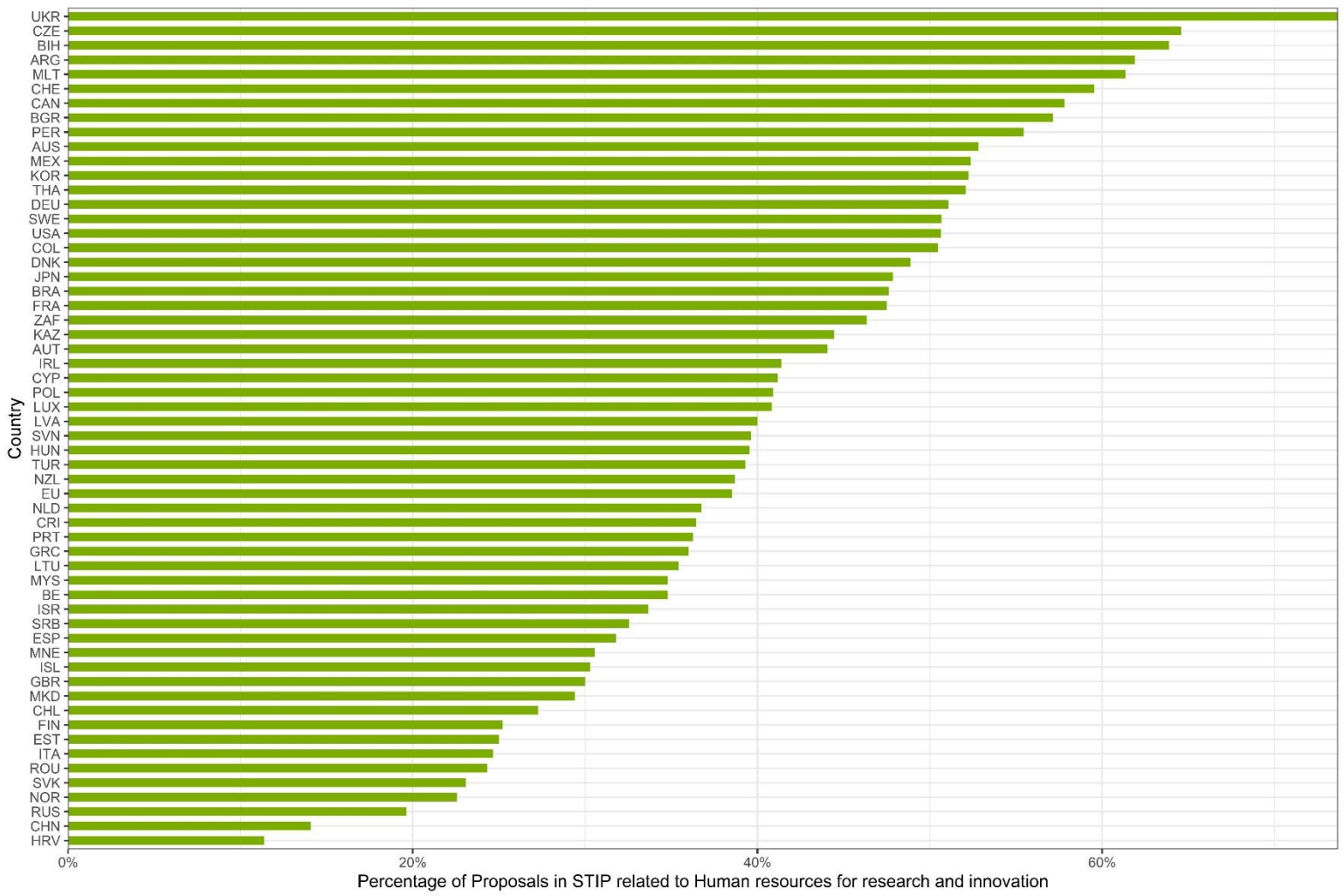
3 Target Groups

- Postdocs and other early-career researchers (TG11, 1556 policies);
- PhD students (TG12, 1207 policies);
- Established researchers (TG9, 1926 policies)

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

Proposals related to HR for R&I per country

57 countries

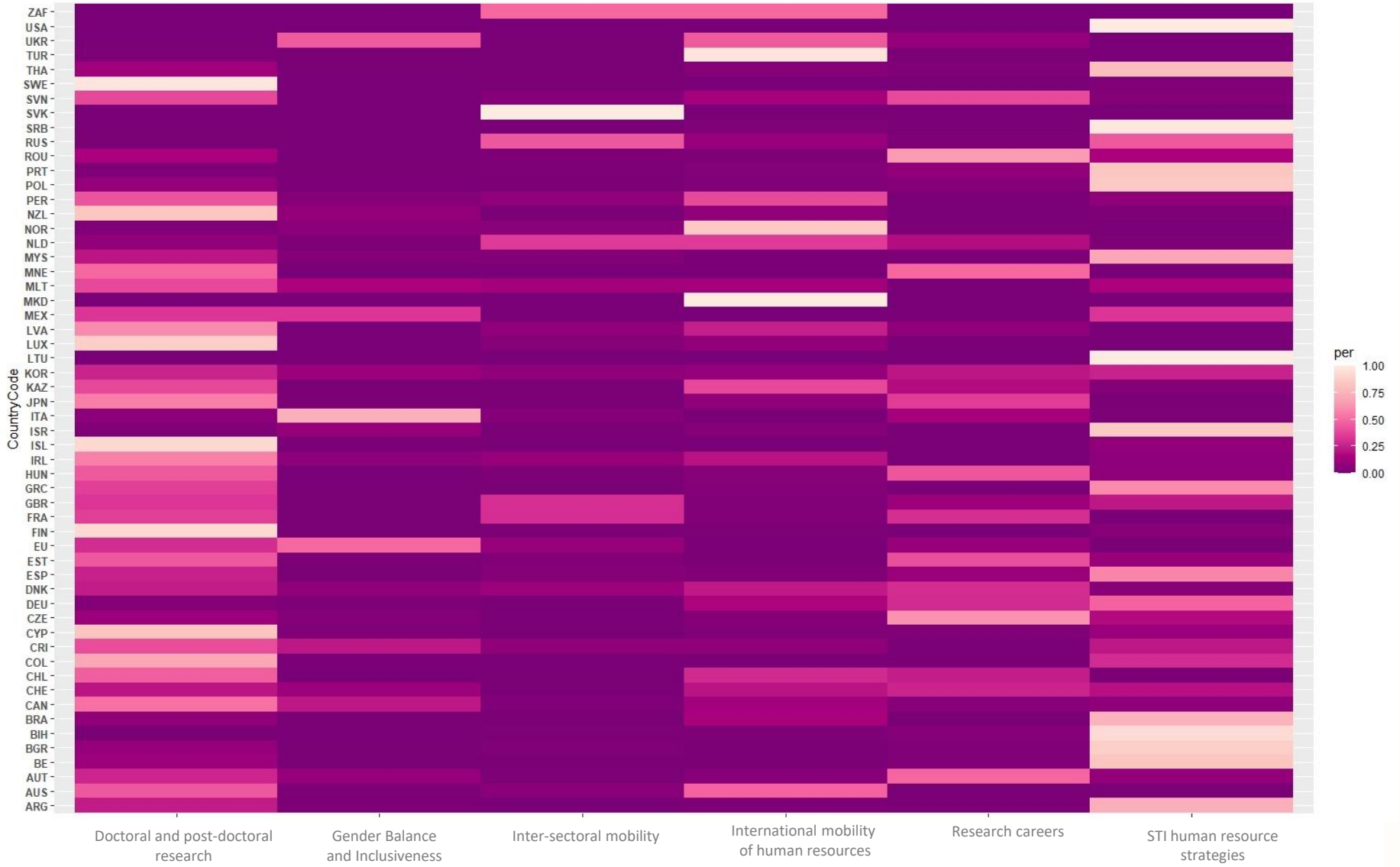


Key Findings

1. Ukraine, Czech Republic and Bosnia Herzegovina most focused on HR (>60%)
2. Croatia, China, Russia less focused on HR (<20%)
3. Most of the countries have between 20% to 60% their STIP proposals focused on Human Resourced for Research and Innovation

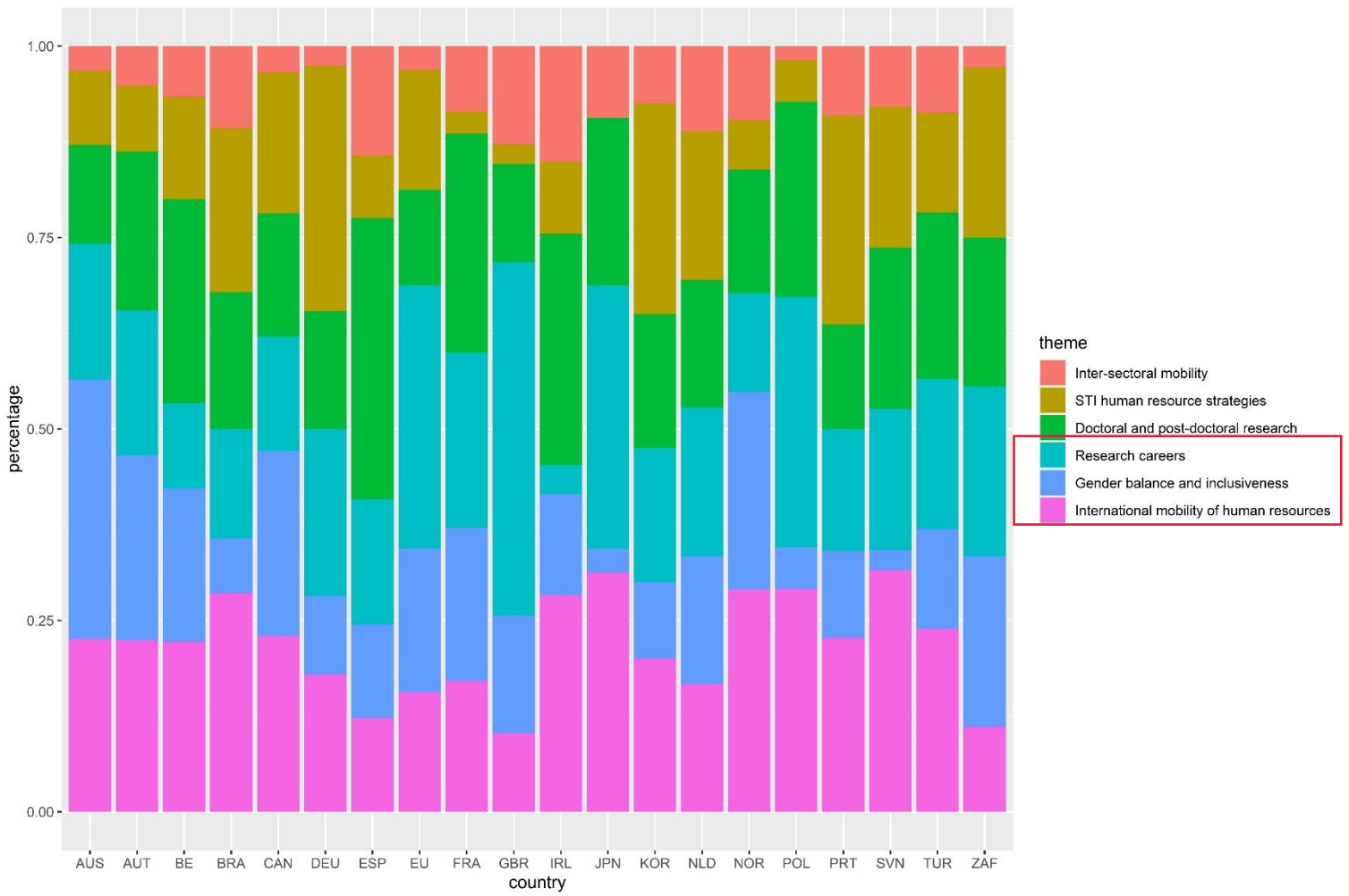
Policy Themes Budget Heatmap

- Percentages are based on distribution of budget per theme within each country
- Clearer colours represent a larger percentage of budget allocated per theme
- Across the majority of countries Doctoral and post-doctoral research and STI human resource strategies have the highest percentage of budget allocation
- Within each country there is high variation of budget percentage allocation
- Some countries are focusing all their budget in one theme, whereas others spread it out



Proposal distribution across themes

Closer look at sample of 20 countries

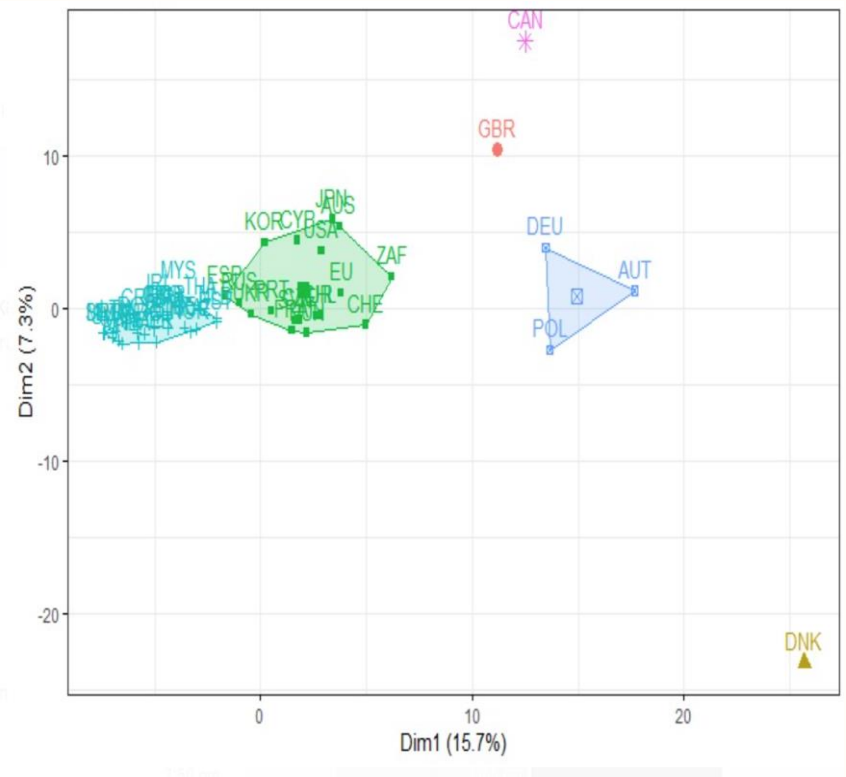


Key Findings

1. Percentages are based on number of policies per theme in each country
2. High variability of theme distribution
3. The analysis will focus on Research Careers, Gender Balance and Inclusiveness and International Mobility of Human Resources

A. Research Careers – Results

Result of PCA



K-means clustering + Manual revise

	Cluster A	*Cluster B*	*Cluster C*	*Cluster D*	*Cluster E*
Key Characteristics	*Employment creation*	*Research capability & Established researchers focus*	*Gender Equality & PostDoc focus*	*Diversification of talents & PhD students focus*	*Others*
Countries	CHE, CHL, CYP, CZE, FRA, HUN, PRT, RUS, SVN, TUR, UKR, ZAF	ARG, BE, BGR, BIH, BRA, CHN, CRI, ESP, EST, MYS, GRC, IRL, ITA, KAZ, LTU, LUX, MEX, MLT, MNE, FIN, NLD, NOR, PER, ROU, SRB, SWE, THA	AUS, DEU, EU, GBR	AUT, POL	JPN, KOR, USA, CAN, DNK
# of countires	12	27	4	2	5
# of policies	79	73	57	29	41

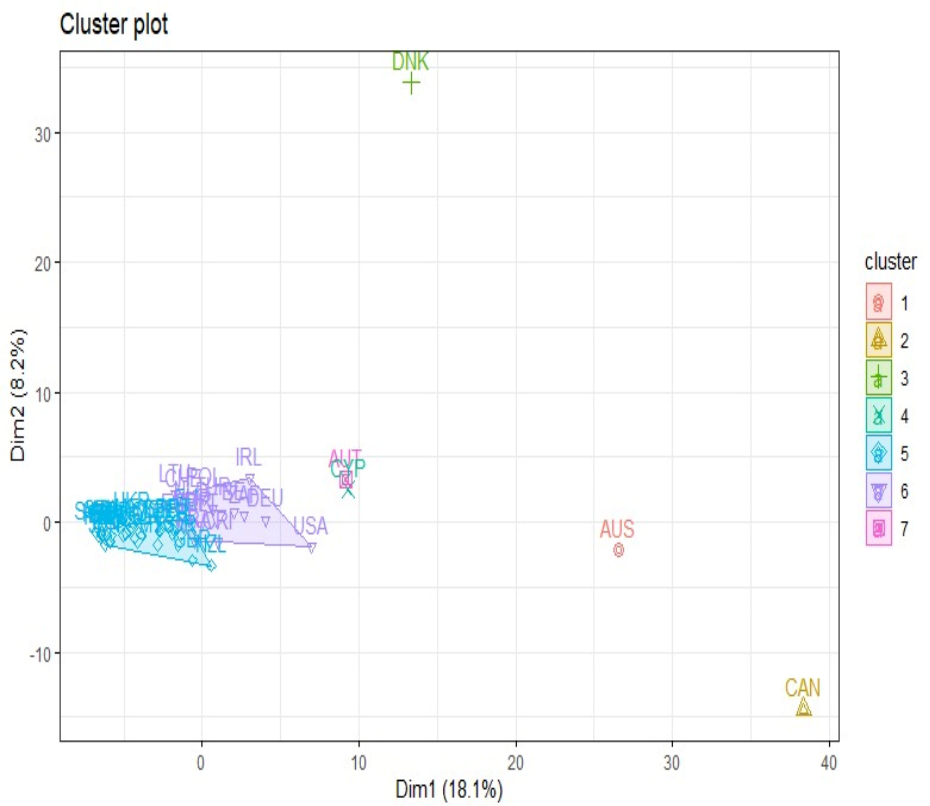
UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

(Reference) Research Careers - Top 10 words

	Cluster A	*Cluster B*	*Cluster C*	*Cluster D*	*Cluster E*
Key Characteristics	***Employment creation***	***Research capability & Established researchers focus***	***Gender Equality & PostDoc focus***	***Diversification of talents & PhD students focus***	***Others***
Top 10 words	scientif	technolog	women	programm	-
	programm	law	equal	doctor	-
	young	scientif	gender	mobil	-
	scientist	sector	organis	erc	-
	focus	innov	charter	conduct	-
	qualifi	number	earli	scientist	-
	employ	phd	stem	scientif	-
	project	public	european	foreign	-
	develop	talent	enabl	opportun	-
	excel	use	career	student	-
Countries	CHE, CHL, CYP, CZE, FRA, HUN, PRT, RUS, SVN, TUR, UKR, ZAF	ARG, BE, BGR, BIH, BRA, CHN, CRI, ESP, EST, MYS, GRC, IRL, ITA, KAZ, LTU, LUX, MEX, MLT, MNE, FIN, NLD, NOR, PER, ROU, SRB, SWE, THA	AUS, DEU, EU, GBR	AUT, POL	JPN, KOR, USA, CAN, DNK
# of countires	12	27	4	2	5
# of policies	79	73	57	29	41

B. Gender Balance and Inclusiveness – Results

Result of PCA



K-means clustering + Manual revise

	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Key Characteristics	*Equality STI policies*	*Discrimination in higher education*	*Diversity in federal universities*	*Under-representation*	*Others*
Countries	ARG, BRA, CHE, CHL, COL, CRI, CZE, ESP, EST, EU, FIN, FRA, GBR, GRC, HRV, HUN, ISL, ISR, ITA, JPN, KAZ, KOR, LTU, MEX, MLT, MYS, NLD, NOR, NZL, PER	BE POL ZAF	AUT DEU IRL USA	CAN	DNK, CYP. AUS
# of Countries	23	3	4	1	3
# of Polices	154	14	37	21	21

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

Gender balance and inclusiveness – Reasoning from literature

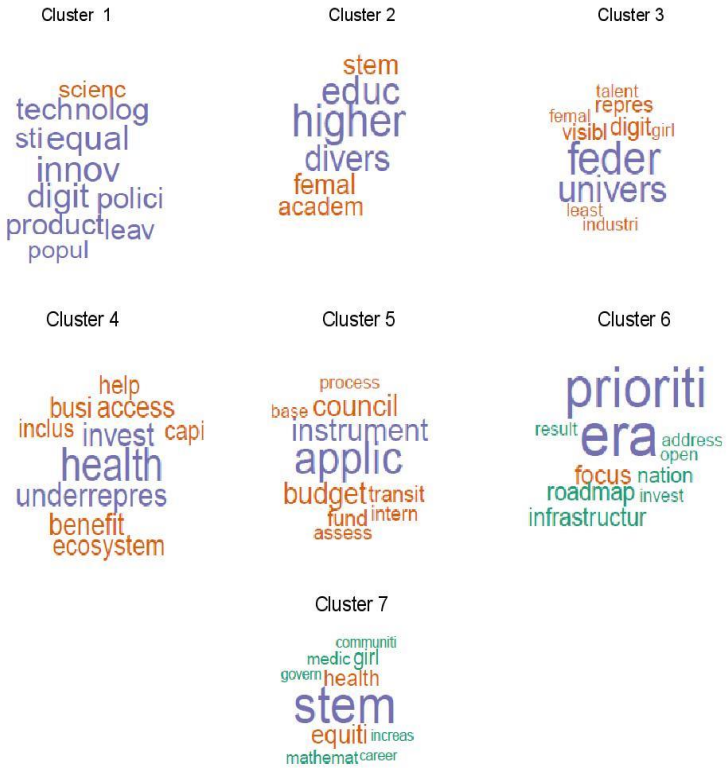
1. The **promotion of equality** between men and women in science, a major policy concern at European level since 1990(Tan et al., 2022).

2. From the analysis, majority of the countries falls under cluster “Equality in STI”

3. One aspect of the **unexplained component of the gender pay gap** is ‘discrimination’, against which most OECD countries have legislated (OECD, 2012). Hence second major typology evolved was 'Assessing discrimination in higher education'.

4. The ‘**Glass ceiling’** for women in reaching **decision making bodies** of science research directly relates to less participation of women in research(Striebing et al., 2020). This explains Cluster 3 & Cluster 4.

(Reference) Gender balance and inclusiveness - Top 10 words

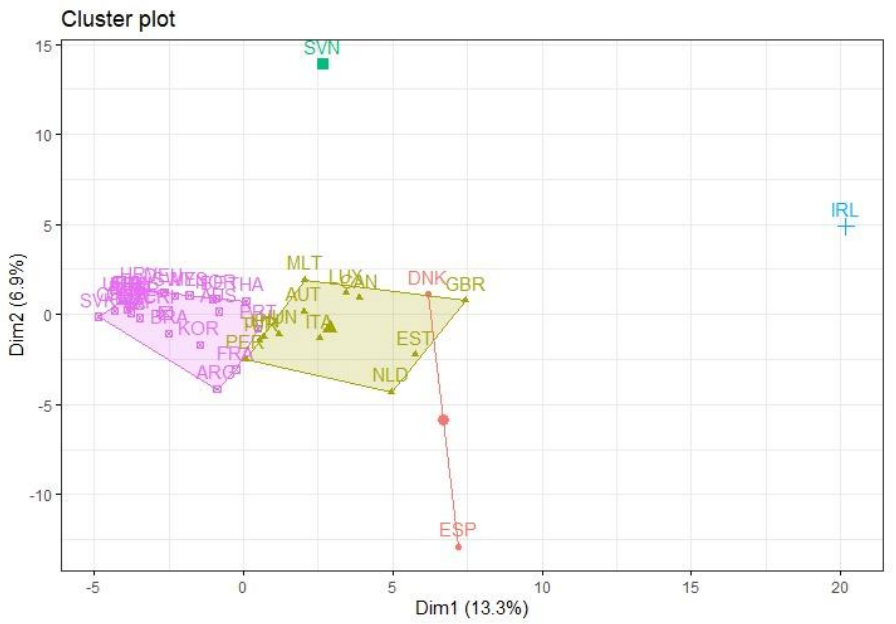


Equality in STI policies	Discrimination in academia	Diversity in federal universities	Under representation	Others		
innovation	academia	federal	underrepresentation	applicant	prioritise	STEM
equal	higher	universities	invest	instrument	infrastruct	equities
technology	assess	education	business	budget	focus	STEM
polices	discrimination	diversity	access	council	roadmap	govern
education	entrepreneurship	increase	benefit	fund	invest	mathematics
digit	number	proportion	capital	Intern	addressing	medic
science	inform	STEM	help		open	increase
higher	issues	visibility	inclusive		result	communities
female	education		ecosystems			girl

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

C. Inter-Sectoral Mobility – Results

Result of PCA



K-means clustering + Manual revise

	Cluster A	*Cluster B*	*Cluster C*
Key Characteristics	*Innovation* *Industry* *Private Research*	*PhD Students* *Established Researchers* *Commercialisation of Public Research*	*Others*
Countries	ARG, AUS, AUT, BE, BGR, BIH, BRA, CHL, CRI, DEU, EU, FIN, FRA, GRC, HRV, HUN, ITA, KOR, LTU, LVA, MYS, NOR, PER, POL, PRT, RUS, SVK, SWE, THA, TUR, UKR, ZAF	CAN, DNK, EST, GBR, JPN, LUX, MLT, NLD	IRL, ESP, SVN
# of Countries	34	8	3
# of Polices	70	27	16

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

(Reference) Research Inter-Sectoral Mobility - Top 10 words

Cluster A

Cluster B

	Cluster A	Cluster B	Cluster C
Key Characteristics	*Innovation* *Industry* *Private Research*	*PhD Students* *Established Researchers* *Commercialisation of Public Research*	*Others*
Top 10 Words	technolog phd innov scientif promot doctor compani industri	student career work opportun partnership skill young collabor	enhanc impact train enterpris peopl educ compani graduat profession
Countries	ARG, AUS, AUT, BE, BGR, BIH, BRA, CHL, CRI, DEU, EU, FIN, FRA, GRC, HRV, HUN, ITA, KOR, LTU, LVA, MYS, NOR, PER, POL, PRT, RUS, SVK, SWE, THA, TUR, UKR, ZAF	CAN, DNK, EST, GBR, JPN, LUX, MLT, NLD	IRL, ESP, SVN
# of Countries	34	8	3
# of Polices	70	27	16

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

Answer to the Research Question

To what extent is it possible to characterise typologies of policy proposals on the theme of scientific employment and research careers?

We were able to characterise typologies under specific themes, for example under the theme research careers five types emerged:

1.Employment creation, 2. Research Capability, 3. Gender Equality and 4. Diversification of Talents, and 5. Others.

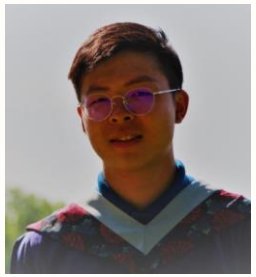
- ✓ **Several typologies were created per each theme studied**
- ✓ **Disaggregation of themes can be explored per cluster of countries**
- ✓ **This could serve as guidance for policy focus on specific areas**
- ✓ **Additionally, clusters within themes can be compared**

Conclusions

- **Text mining** such as hierarchical clustering, k-mean clustering and PCA can be applied as tools for analysing policies and extracting information beyond the survey answers.
- We can explore **more granular typologies within themes** from different lenses.
- Powerful methodology, however **difficult to interpret** and must be complemented with **additional research to triangulate** the mechanically produced clusters
- Algorithms are sensitive which can lead the methodology to create different results.
- **Unbalanced datasets**, for example, distribution per country can affect the direction of the results.

Thank you – The SPRU team

Anas Aleassa
MSc Sustainable Development



Yongyuan Huang
PhD Science and Policy Studies

Daniela Valenzuela
MSc Strategic Innovation Management

Saradha Krishnamoorthy
MSc Sustainable Development

Satoshi Shimayoshi
MSc Science and Technology Policy

Jiyoung Park
MSc Science and Technology Policy

Jongho Jung
MSc Science and Technology Policy

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

References

- *Auriol, L., Misu, M., & Freeman, R. A. (2013). Careers of doctorate holders: Analysis of labour market and mobility indicators.*
- *Ding, C., & He, X. (2004). K-means clustering via principal component analysis. Proceedings of the twenty-first international conference on Machine learning,*
- *e Conhecimento, R. d. I. (2017). OECD Science, Technology and Innovation Outlook 2016.*
- *OECD. (2012). Closing the Gender Gap: Act Now. OECD. https://doi.org/10.1787/9789264179370-en*
- *Striebing, C., Kalpazidou Schmidt, E., Palmén, R., Holzinger, F., & Nagy, B. (2020). Women Underrepresentation in R&I: A Sector Program Assessment of the Contribution of Gender Equality Policies in Research and Innovation. Evaluation and Program Planning, 79, 101749. https://doi.org/10.1016/j.evalprogplan.2019.101749*
- *Tan, M., Saglamer, G., & Celik, O. (2022). Women in Science, Engineering and Technology (SET) in Mediterranean Basin.*