Who Do You Think You Are?

Identifying Research Software Engineering Personas From Developer / Repository Interaction Data

Felicity 'Flic' Anderson[†⋆], **Dr. Julien Sindt**[†] **& Prof. Neil Chue Hong**[†]

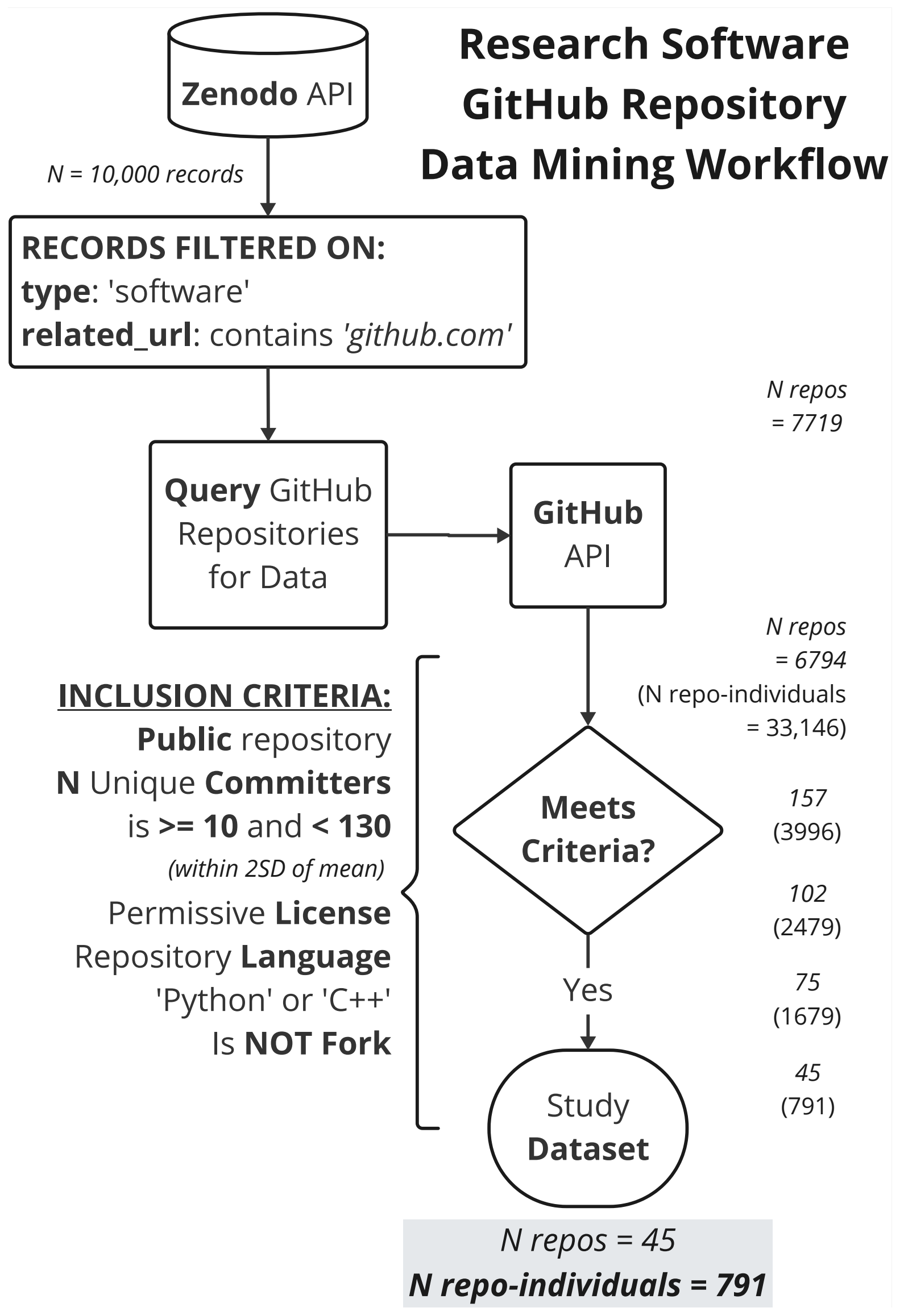
[†]EPCC, University of Edinburgh [⋆]Felicity.Anderson@ed.ac.uk

0: RSE Personas from GitHub Data

We mined 45 GitHub repositories to attempt to identify **data-driven RSE Personas** for 791 contributors.

We hypothesised 4 main RSE Personas (Fig.2); we can confirm **'Active Leaders'** and **'Occasional Contributors'**, others are less clear...

1: How to Create Data-Driven RSE Personas



Research Software GitHub Repository Data Mining Workflow

Zenodo API

N = 10,000 records

RECORDS FILTERED ON:
type: 'software'
related_url: contains *'github.com'*

N repos = 7719

Query GitHub Repositories for Data

GitHub API

N repos = 6794
(N repo-individuals = 33,146)

INCLUSION CRITERIA:
Public repository
N Unique **Committers** is **>= 10** and **< 130** *(within 2SD of mean)*
Permissive **License**
Repository **Language** 'Python' or 'C++'
Is **NOT Fork**

Meets Criteria?

157 (3996)
102 (2479)
75 (1679)
45 (791)

Yes

Study **Dataset**

N repos = 45
N repo-individuals = 791

Fig. 1: Data gathering workflow.

Zenodo was used to source for GitHub repositories with a DOI.

Hypothesised RSE Personas

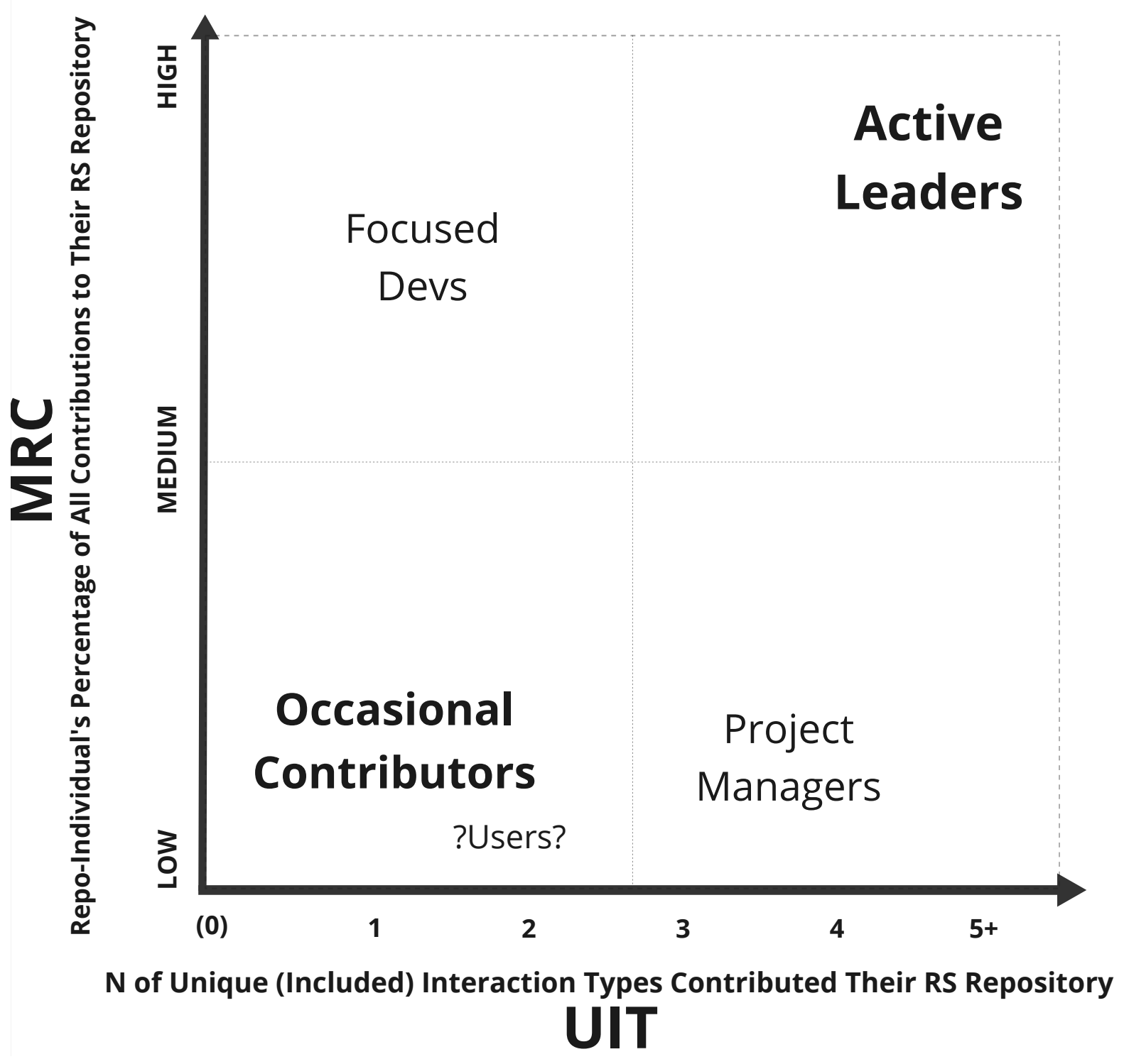
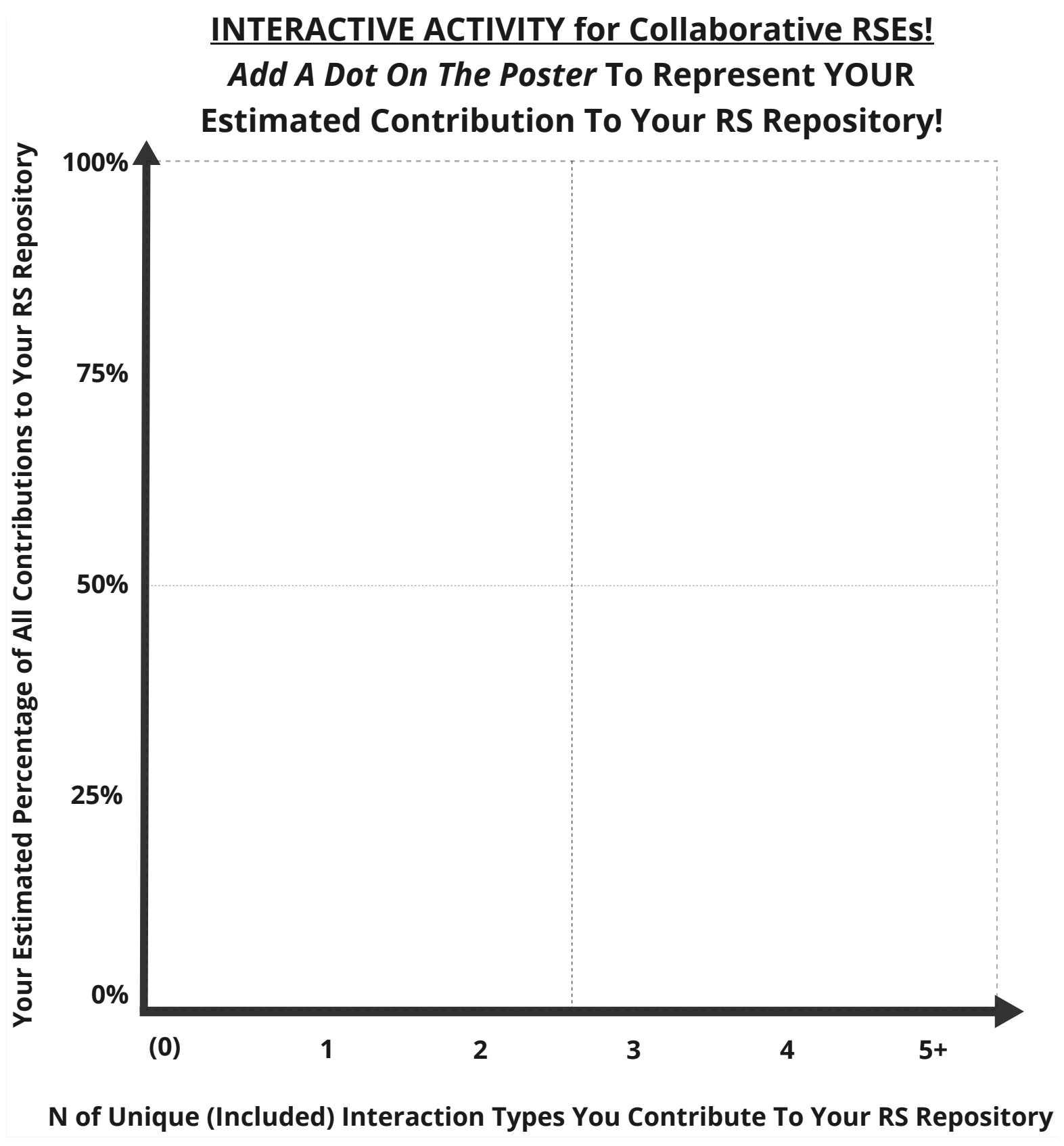

MRC
Repo-Individual's Percentage of All Contributions to Their RS Repository
HIGH
MEDIUM
LOW
Focused Devs
Active Leaders
Occasional Contributors
?Users?
Project Managers
(0)
1
2
3
4
5+
N of Unique (Included) Interaction Types Contributed Their RS Repository
UIT

Fig. 2: Hypothesised RSE Personas.

Repositories in the dataset contained a **mean of 18 repo-individuals** (min=10, max=50). Repository interactions were analysed across variables (see Fig.5)



INTERACTIVE ACTIVITY for Collaborative RSEs!
Add A Dot On The Poster **To Represent YOUR Estimated Contribution To Your RS Repository!**

Your Estimated Percentage of All Contributions to Your RS Repository
100%
75%
50%
25%
0%
(0)
1
2
3
4
5+
N of Unique (Included) Interaction Types You Contribute To Your RS Repository

approximating *'depth' (MRC)* and *'breadth' (UIT)* of contributions.

Hierarchical Clustering identified **3 clusters** (Fig.4: Calinski–Harabasz index 757.60, N=3 clusters) from similarities across interaction data.

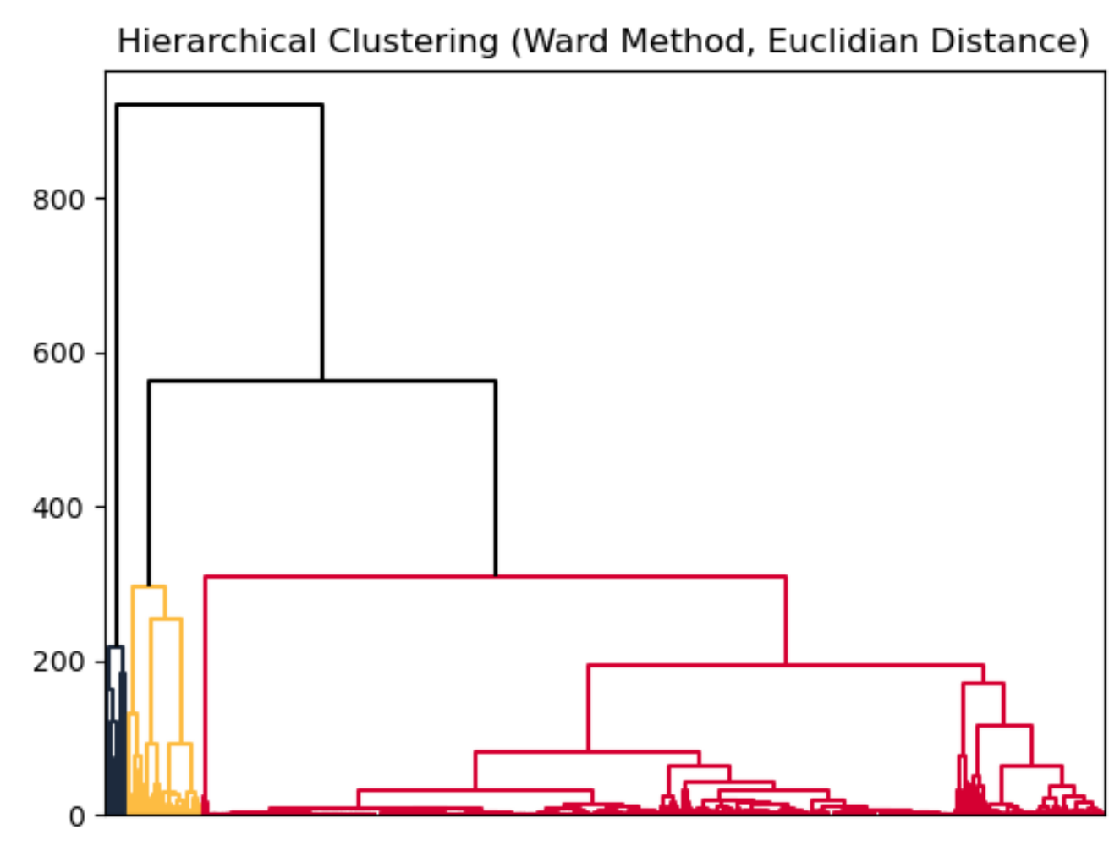


Fig. 4: Dendrogram for Hierarchical Clustering using Ward method and Euclidean distance.

714 repo-individuals fall into one cluster category, with the remaining 77 across two remaining categories. **Cluster 0: 90.27%**, (red); **Cluster 2: 7.46%**, (gold); **Cluster 1: 2.28%**, (navy).

Cluster diversity was moderate: 84.4% (38) of the 45 repositories contain RSEs from 2 clusters. 6 (13.3%) had all clusters, 1 repository contained only a single cluster.

VARIABLE	DESCRIPTION & CALCULATION	CONCEPT
Repo-Individual	**An RSE**, contributing 1+ interaction within a RS repository in the study. **Repo-Individual = GH Username + Repository Name combination**	The unit of study.
Included Interaction Types	Five **types of repository interactions** covered by this study. ***"commit created", "issue ticket created", "issue ticket closed", "issue ticket assigned to individual", "pull request created"***	Interaction Types examined to generate RSE Personas.
Unique Interaction Types	**UIT = Unique categories of interaction** Repo-Individual has made within their repository. (Minimum = 1, Maximum = 5). **UIT = +1 for each Unique Included Interaction Type**	BREADTH of repository interactions all included types.
Repository Contribution ***[Interaction Type]***	**RC[type] = Percentage of repo's interactions of [type]** by this Repo-Individual. **% RC[type] = (Repo-Individual's Number of Interactions** of [Interaction Type] **/ Total Interactions** of *[Interaction Type]* **for their Repository) * 100** E.g. Repo has 1000 commits, and a repo-individual created 200: (200 / 1000) * 100 = 20%	DEPTH of repository interactions *of a specific type.*
Mean Repository Contribution	**MRC = Mean percentage score across ALL Interaction Types** for a Repo-Individual. **% MRC = (RC1 + RC2 + RC3 + RC4 + RC5) / Max UIT (5)** E.g. Repo-individual's score might be (200/1000 + 10/50 + 5/15 + 0/5 + 1/2) / 5 = 24.6%	DEPTH of repository interactions of *all included types.*

Fig. 5: Key variable definitions and relationship to hypotheses.

2: Identifying (with) RSE Personas

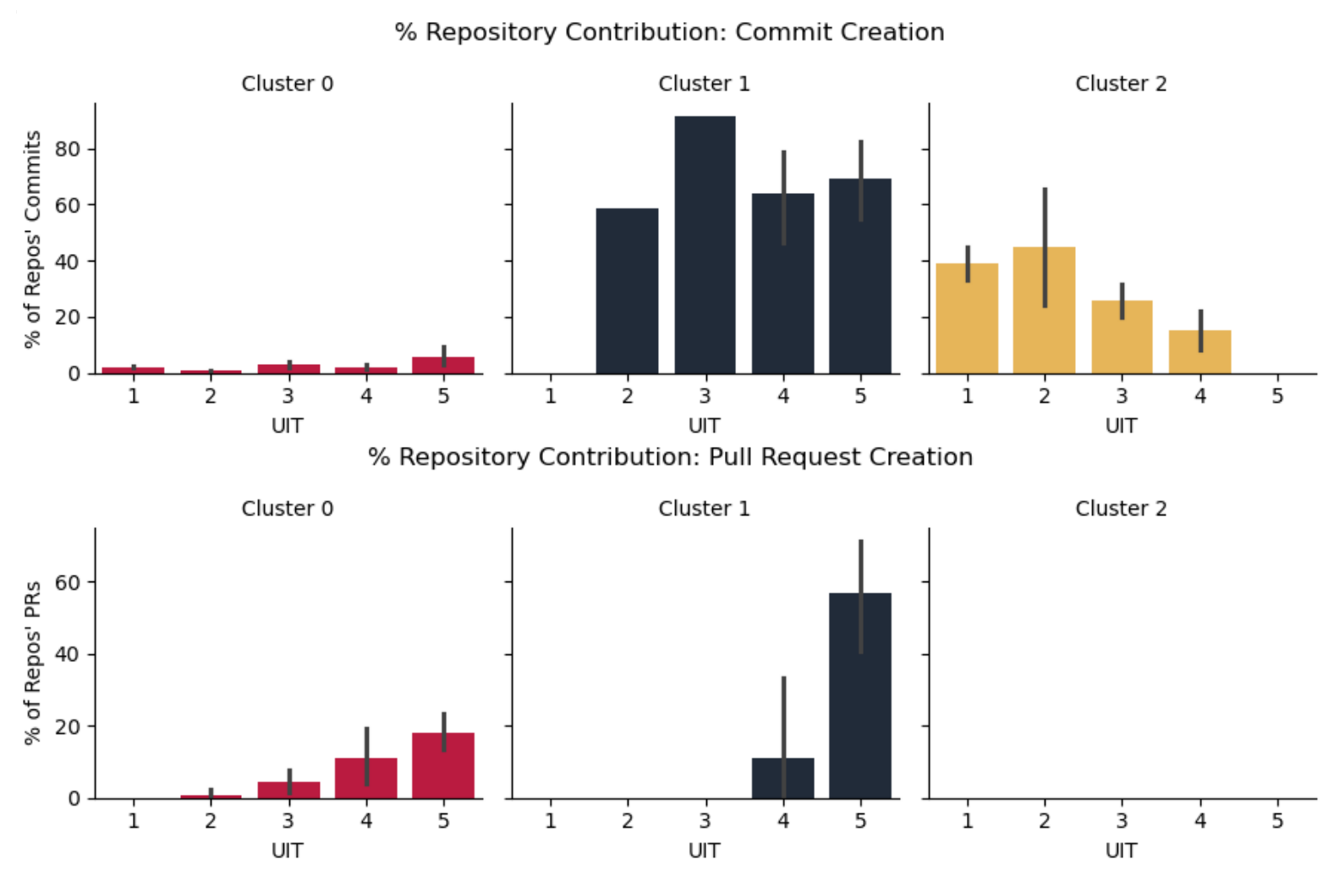


Fig. 6: RC values for Commit Creation and Pull (PR) Request Creation, across UIT.

Fig.6 shows contrasting patterns of behaviours between clusters for committing and PRs, supported by ANOVA results (p < 0.001).

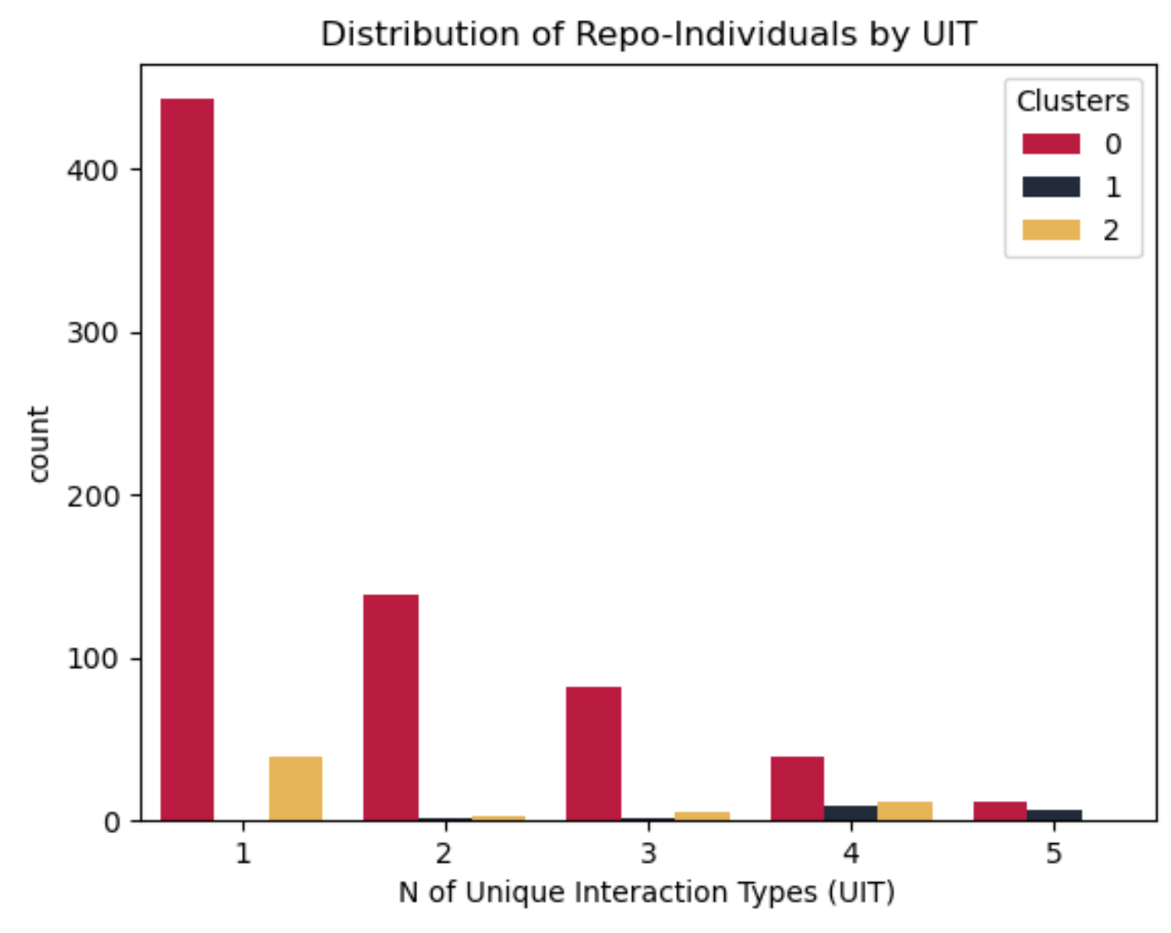


Fig. 7: Distribution of Repo-Individuals across UIT.

Fig.7 indicates the majority of repo-individuals only contribute to 1 UIT. These are the 482 (60.9%) RSEs from 41 repositories who only create commits. Conversely, only 18 (2.3%) repo-individuals displayed interactions from all valid interaction types.

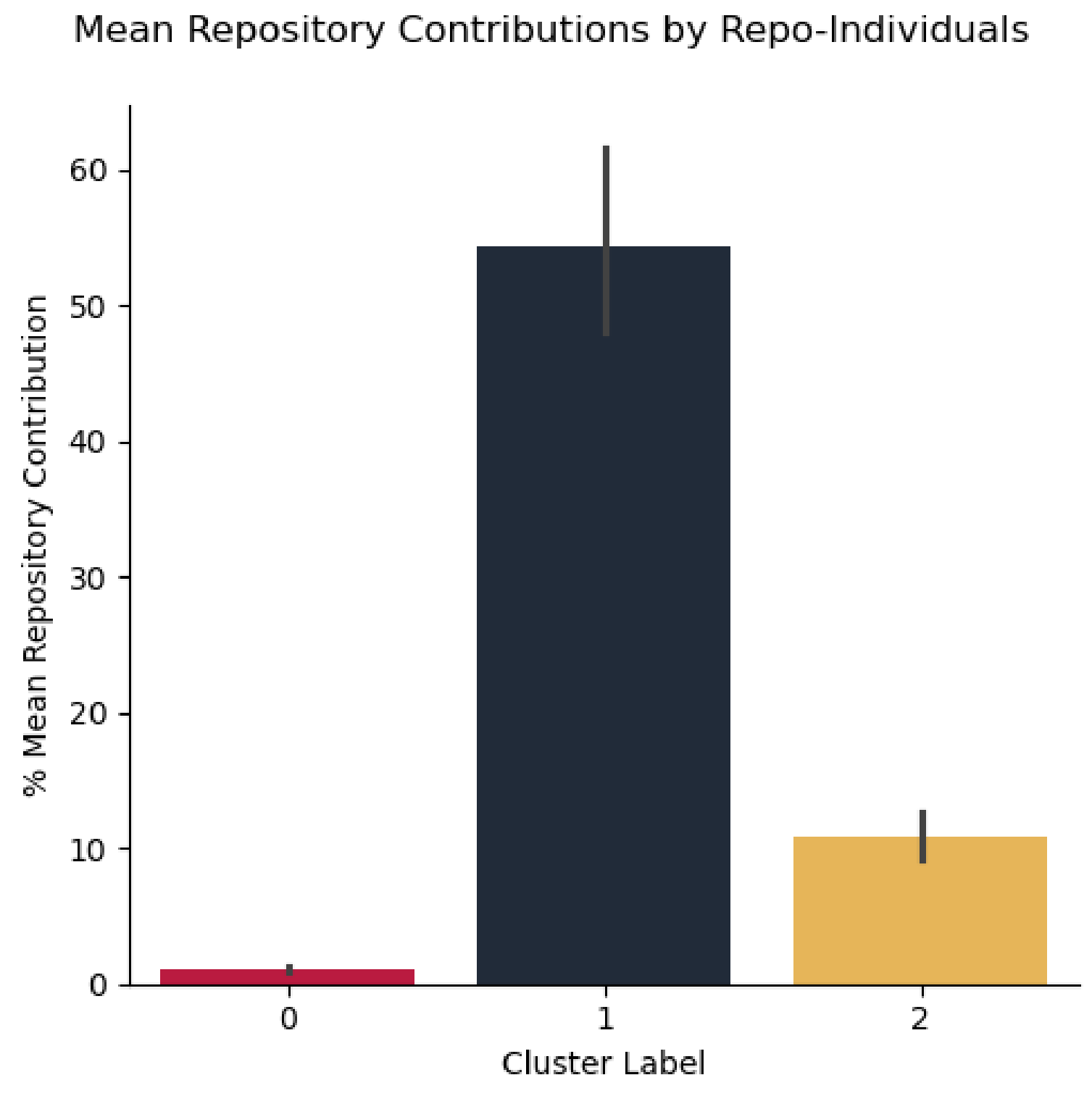


Fig. 8: MRC highly significant (p < 0.001) by Cluster.

Differences in MRC between clusters are highly significant (p < 0.001) across all relationships (Fig.8).

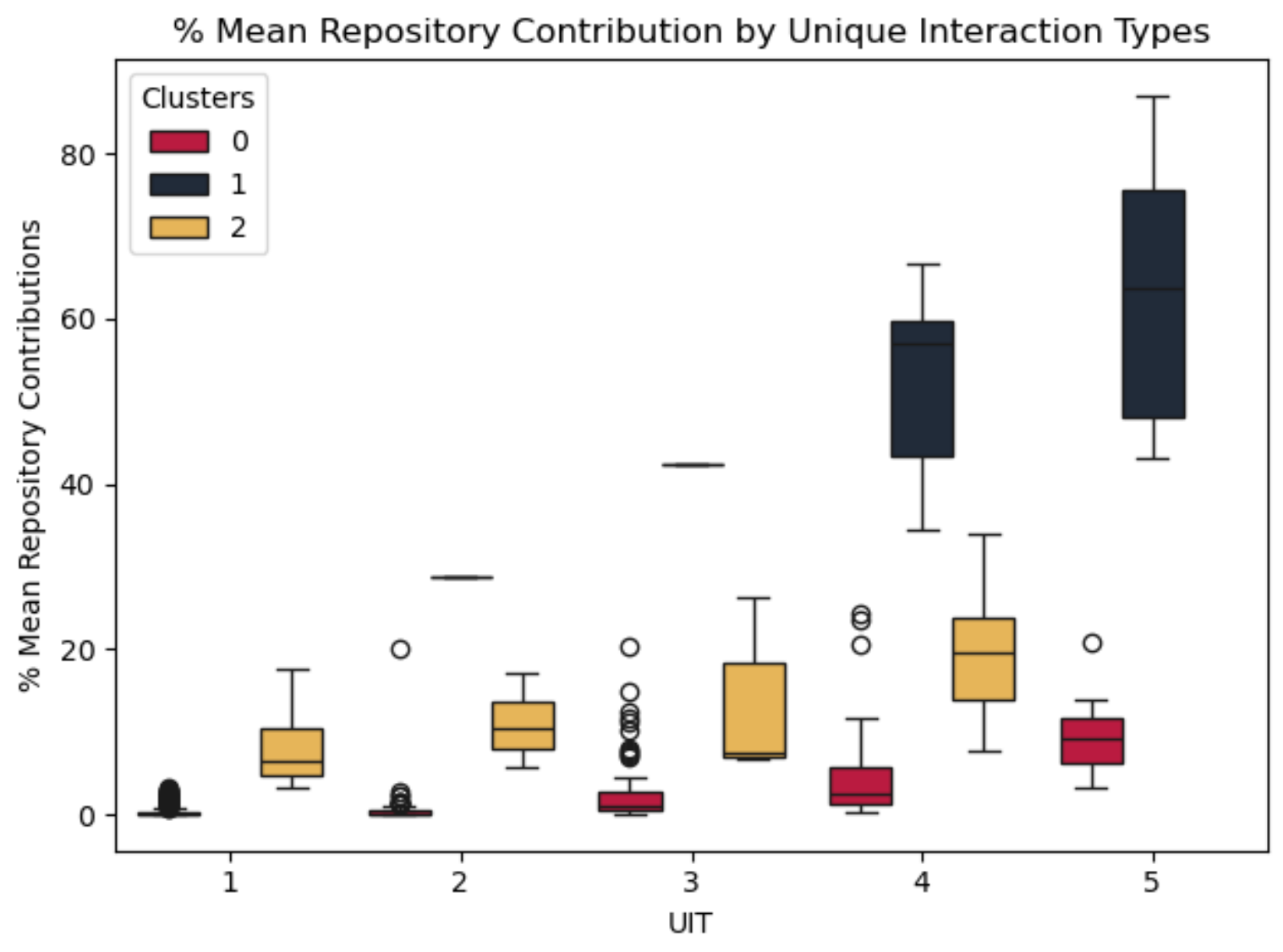


Fig. 9: Variation in % MRC across UIT.

ANOVA and Tukeys HSD tests identify MRC varying significantly across UIT categories (p < 0.001), while low:high pairings of UIT are highly significant.

3: Evaluating RSE Personas Approach

Active Leaders map to Cluster 1 as a RSE Persona type with **high UIT and high MRC** values (Fig.9). Despite rarity (2% of repo-individuals), they generate >50% their repositories' interactions.

Cluster 0 maps well to hypothesised 'Occasional Contributors', showing low 'MRC' values and skewing strongly to very low 'UIT'. Could be examined for possible sub-clusters.

'Focused Developers' RSE Persona was disproved by no clusters with high MRC and low UIT, so is rejected as hypothesised.

The **'Project Managers'** RSE Persona is rejected as unclear - Cluster 2 does occupy the low MRC space, but may show possible bimodality in relation to UIT.

Variables 'MRC' and 'RC pull-request-created' are significant and explain high variance in Principal Component Analysis (PCA), acting as a robust basis for RSE Personas. **UIT does not explain cluster variance well.** This could be due to the differing 'combinations' of interactions confusing the picture.

Ask Me About Limitations... Skews towards 'best practices'; not all interactions are equal; project comparison difficulties; capturing changes over time; real-world validity-checking...

4: Using RSE Personas

Next Steps... Include more UITs (code review, discussions, merges); compare other languages - data vs HPC?; explore time and commit category data more deeply...

Future Work... *RSE Persona dynamics* - co-occurrences or limiting factors?; *RSE Persona Fluidity* - do personas change over time/repos?); *create Taxonomy of RSE Personas* with descriptive characteristics to aid usage.

github.com/FlicAnderson/RSE-Personas-Who |

Acknowledgements: Work funded by EPSRC DTA Scholarship.