
Who Do You Think You Are?

*Creating **RSE Personas** from GitHub Interactions in Mid-Size Research Software (RS) Repositories*

***Flic Anderson**, Dr Julien Sindt, Prof Neil Chue Hong
(EPCC, University of Edinburgh)*

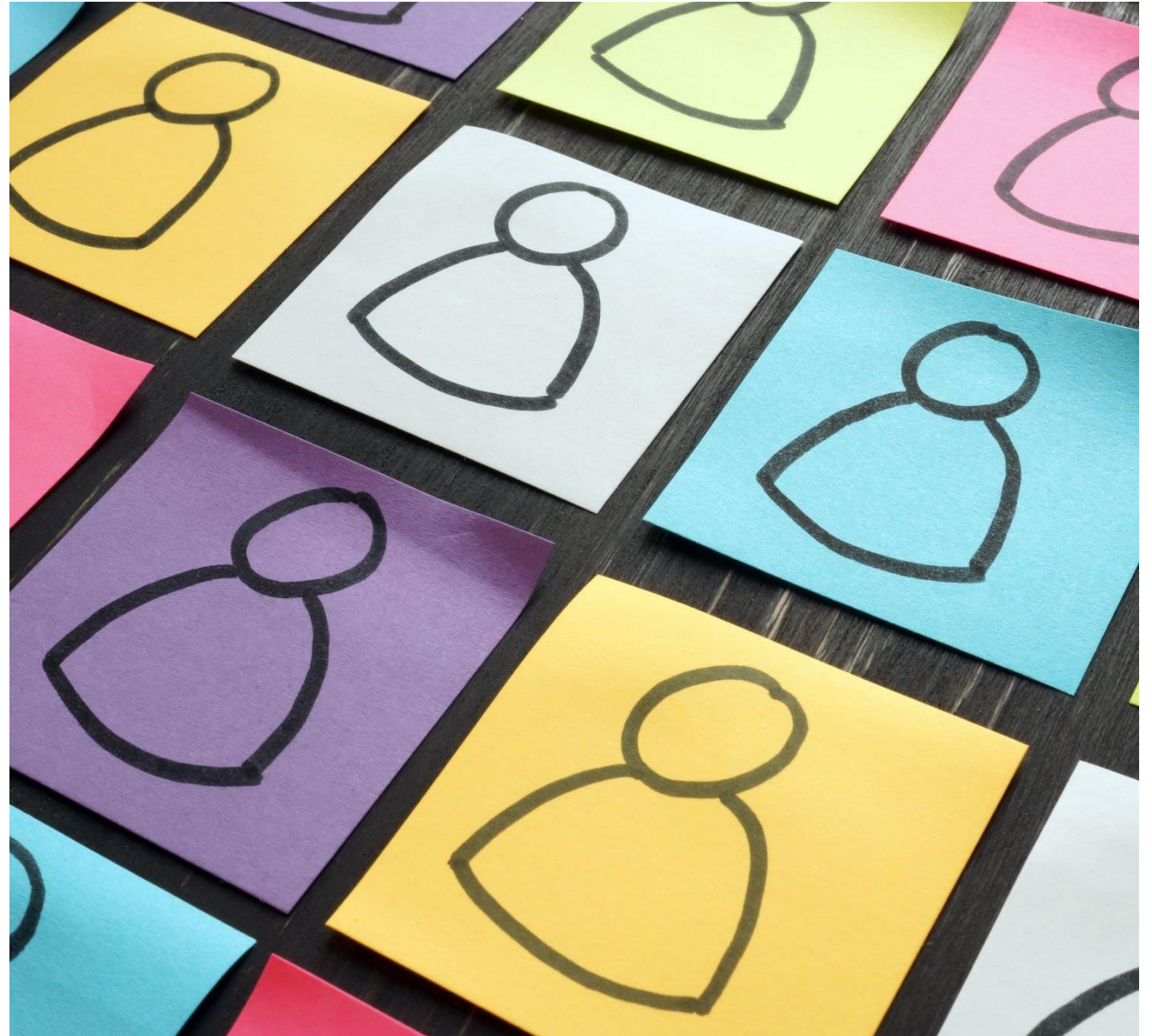
Preprint: doi.org/10.48550/arXiv.2510.05390

RSE Personas:

Patterns of collaborative
Research Software (RS)
Repository interactions on
GitHub.

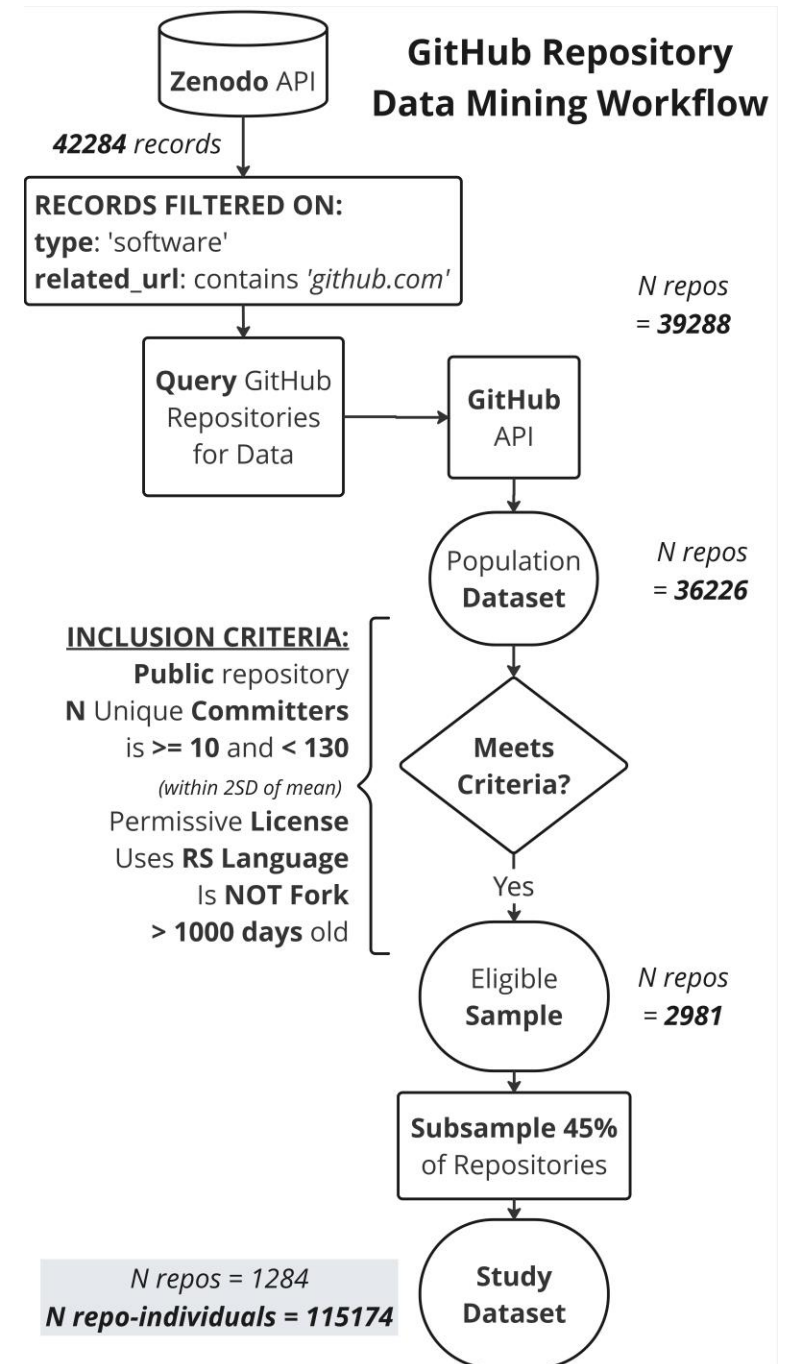
Why?

If we can *name something*,
we can **think, talk about,**
and **change** it!



Research Objectives:

- **Re-find initial clusters from pilot study** (High and Low Interactivity)...
- Applying further clustering to **identify additional high-resolution personas** within these low-resolution groupings
- Proving whether **high-responsibility type interactions** (Issue Assignment, PR Closure) strongly influence RSE Persona creation



Pilot Study Clusters

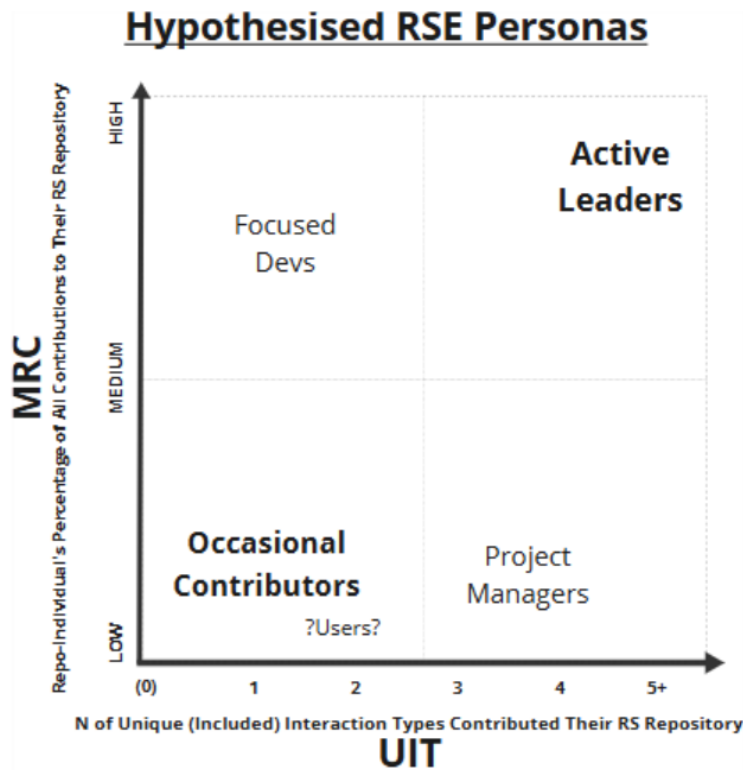


Fig. 2: Hypothesised RSE Personas.

- [Pilot study](#) exploring where and how I might find personas... (45 repos, 791 repo-individuals)
- Originally expected personas in 4 quadrants
- **UIT**: Unique Interaction Types
 - Interaction **Variety**
- **MRC**: Mean Repository Contribution
 - Interaction **Volume**
- **Low to High Axes**
- **Found good evidence for 2...**

Data Analysis:

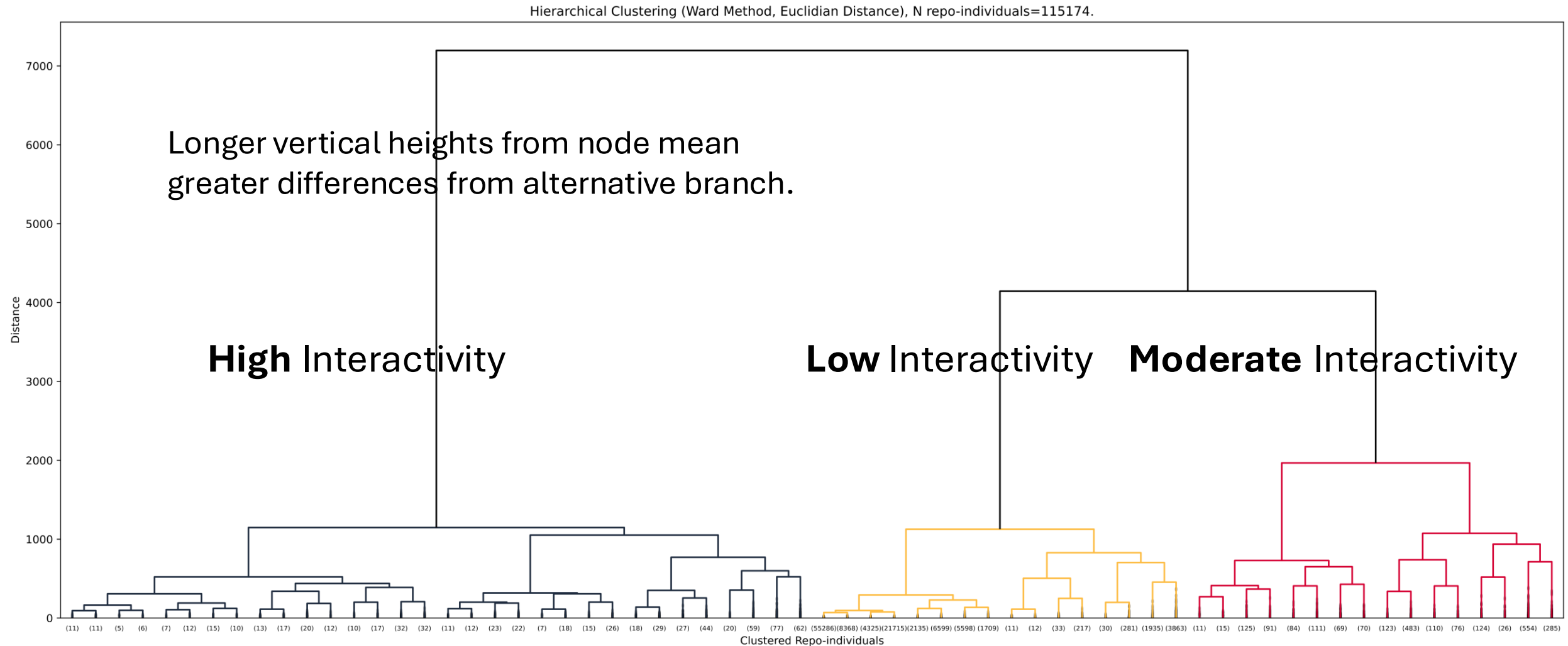
Per-Repo-Individual data on 6 Interaction Types

*Variables focus on **Variety** and **Volume** of interactions*

*Initial clustering / analysis, then **re-cluster subsets** for **better resolution***

Variable	Description: "The Percentage (%) of..."
RC Commit Created	... repository's commits created by repo-individual
RC Issues Created	... repository's issue tickets created by this repo-individual
RC Issues Closed	... repository's closed issue tickets which were closed by repo-individual
RC Issues Assigned of Assigned	... repository's assigned issue tickets which were assigned to this repo-individual
RC Pull Request Created	... repository's pull requests created by repo-individual
RC Pull Request Closed	... repository's pull requests closed by repo-individual
MRC: Mean Repository Contribution	... contribution across all repository interactions for this repo-individual. RC values for all interaction types are summed, and divided by the number of included interaction types to obtain a meta-average, showing the typical interactivity level of this repo-individual. If no contributions made for an interaction type zeroes are used; divisor is always 6 - the number of interaction types
Percentage Created-Closed Issues	... RC of Issues Created by this repo-individual minus RC of Issues Closed by them; indicates whether they are a net creator or closer of issues.
Percentage Sum N Interactions	... repository's total interactions which were contributed by this repo-individual
Percentage Interaction Days	... the Sum of Interaction Days of all repo-individuals for this repository, indicating the proportion of time contributed by this repo-individual compared to other contributors.

Hierarchical Clustering generated 3 initial clusters based on Interactivity Groupings



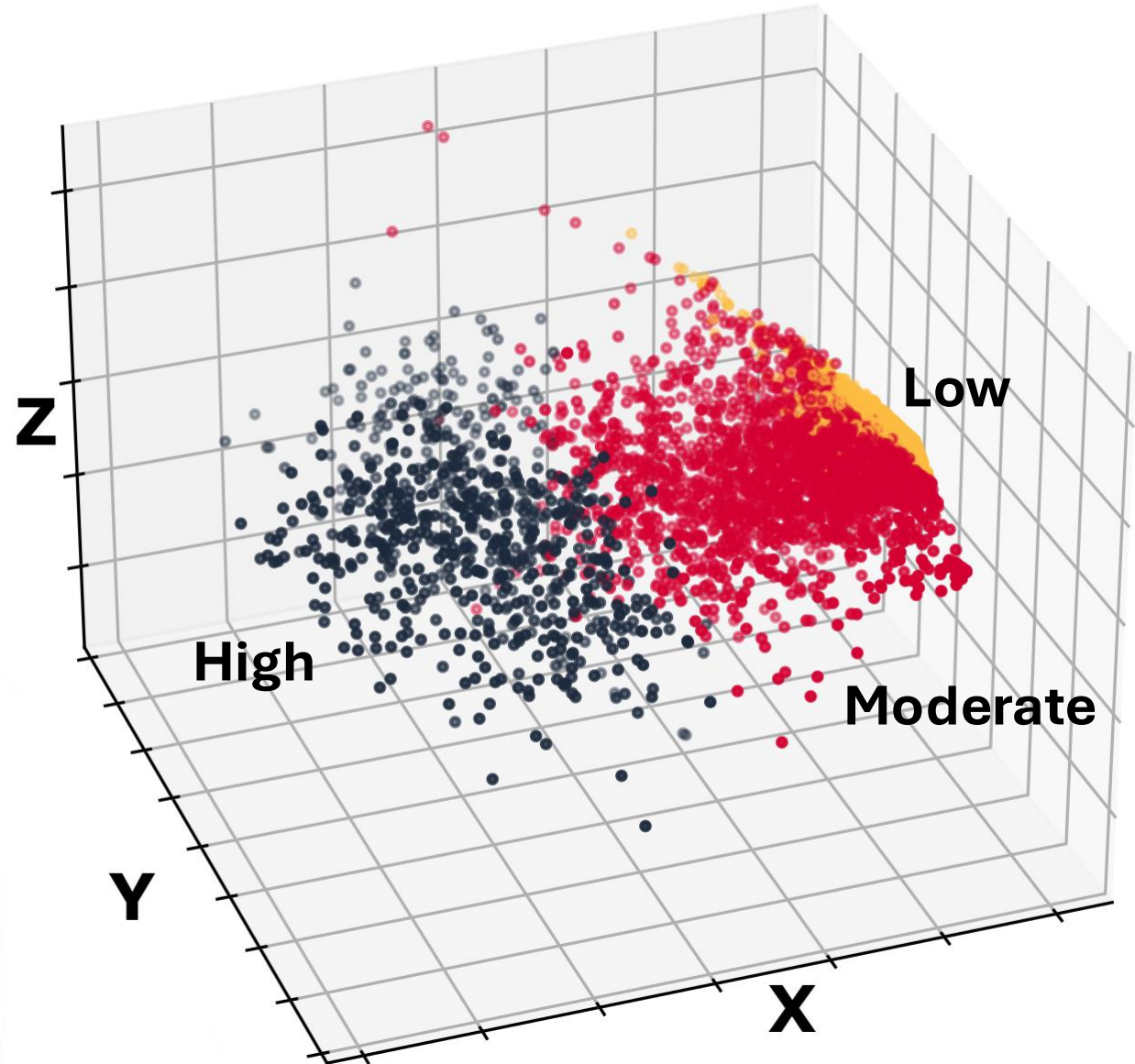
Principal Component Analysis

PCA Eigenvectors explain:

X: 81.36%

Y: 5.54% ...

Z: 4.58% of variance in data



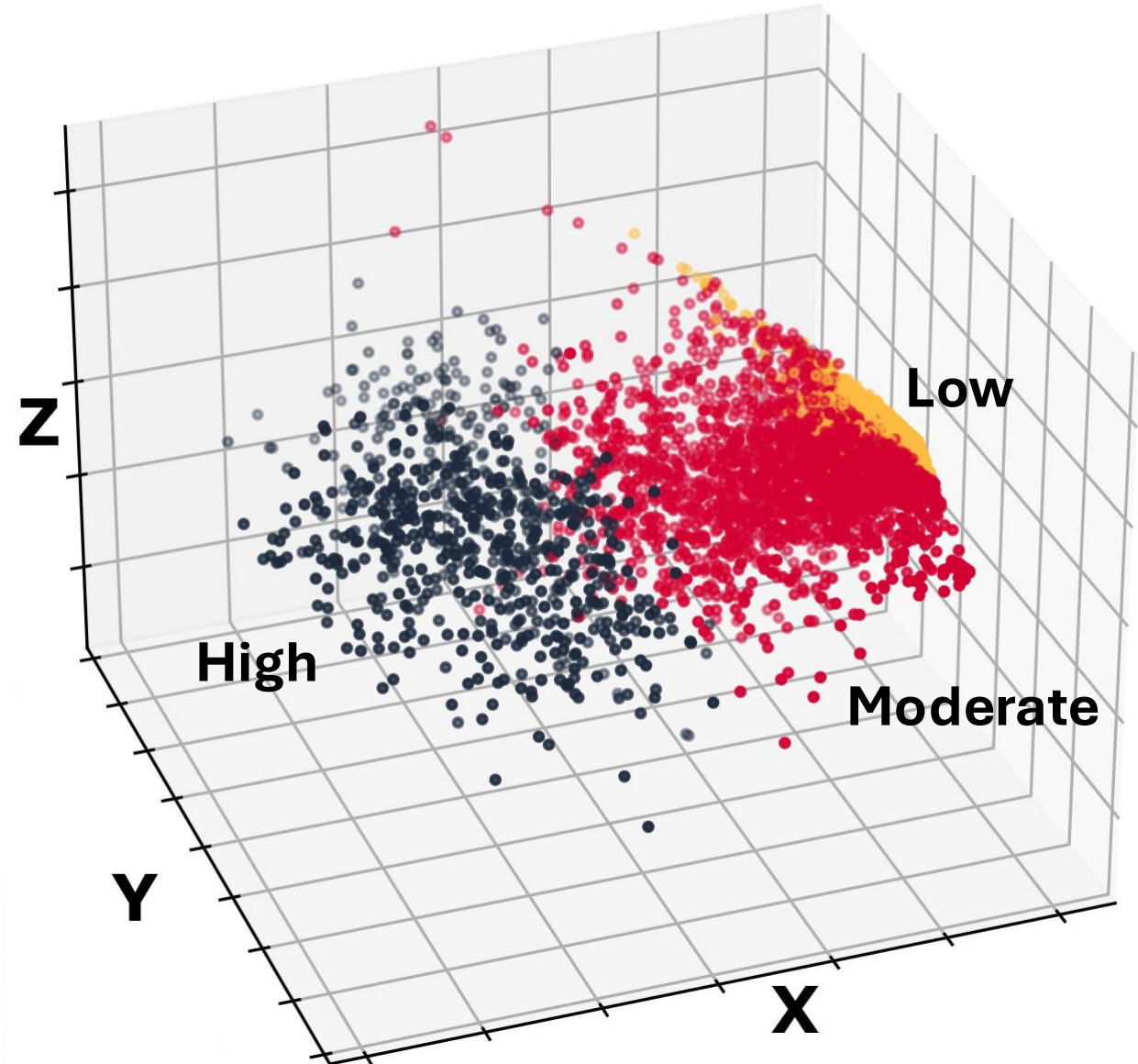
Feature Importance Analysis

Highest Importance Values:

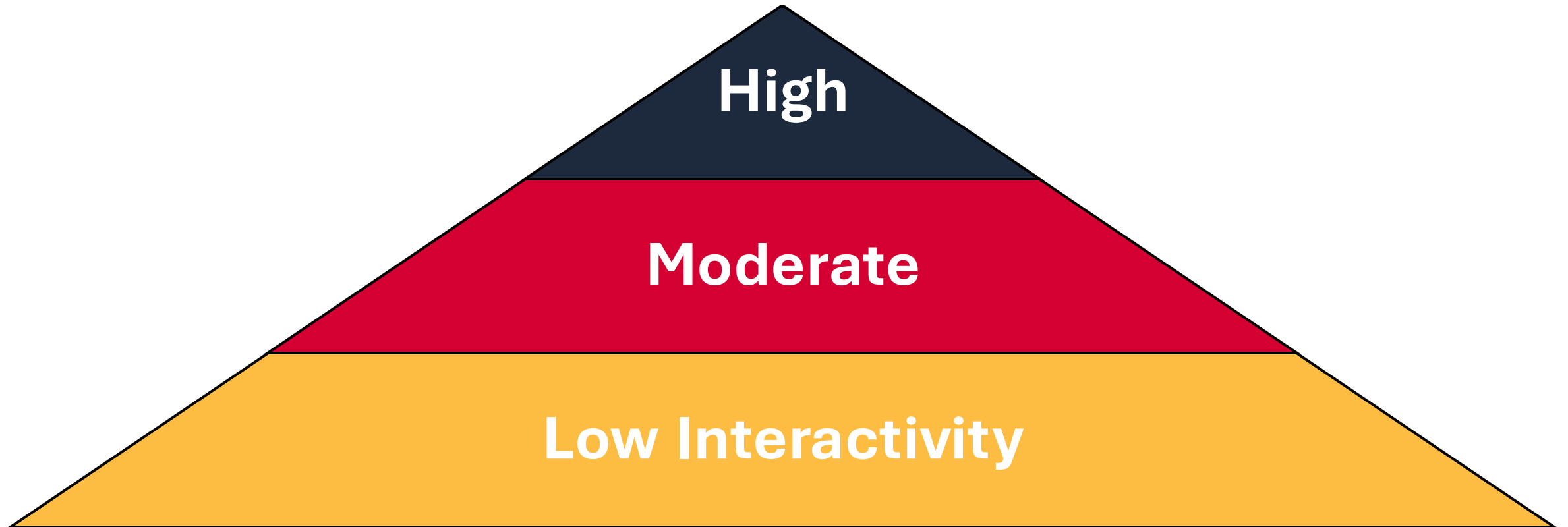
X: RC PR Closed (39.81)

Y: RC Commit Created (27.74)

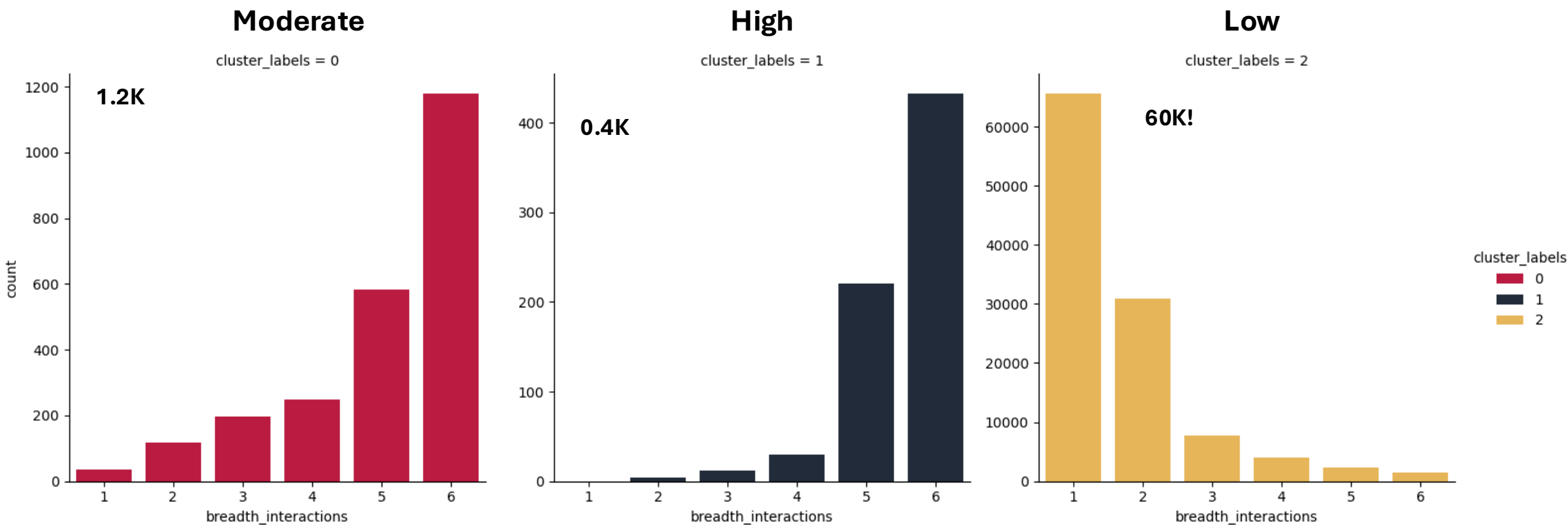
Z: RC Issue Closed (46.08)



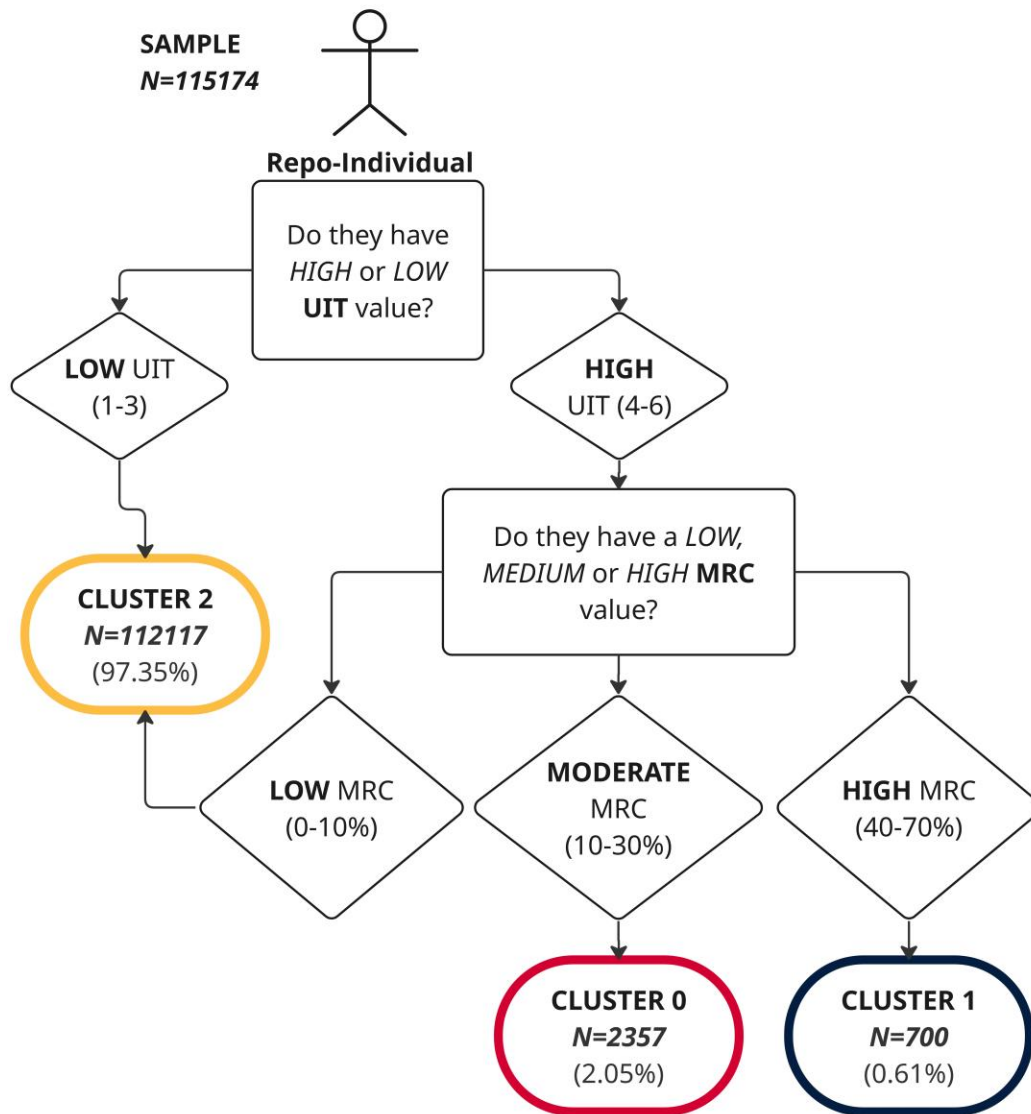
3 levels of general interactivity groupings (considering both variety and volume)



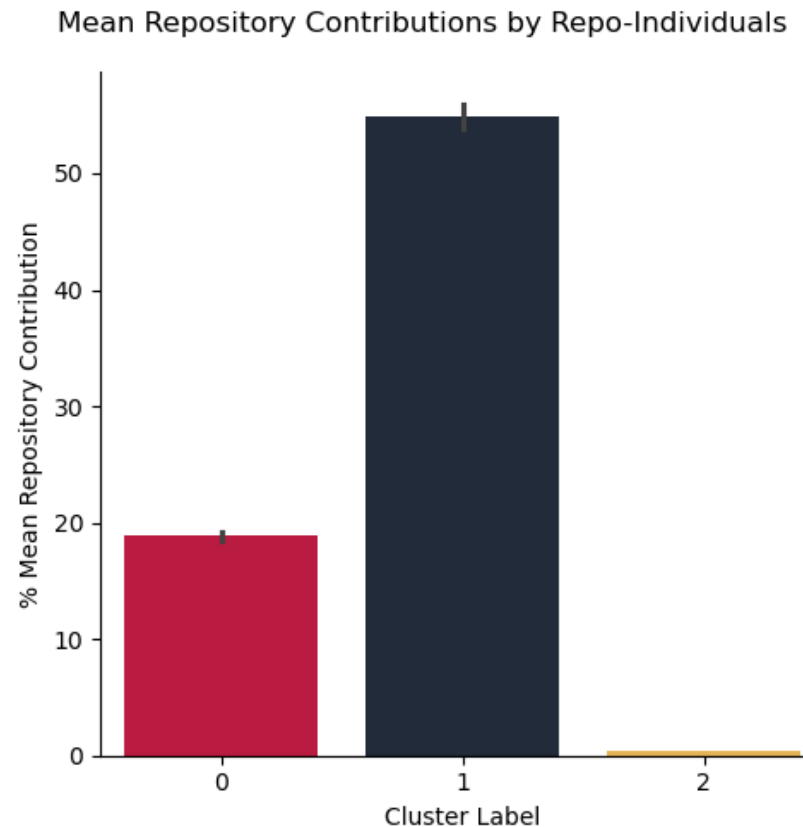
Understanding the Initial Clusters: UIT



Unique Interaction Types (UIT)



Understanding the Initial Clusters: MRC



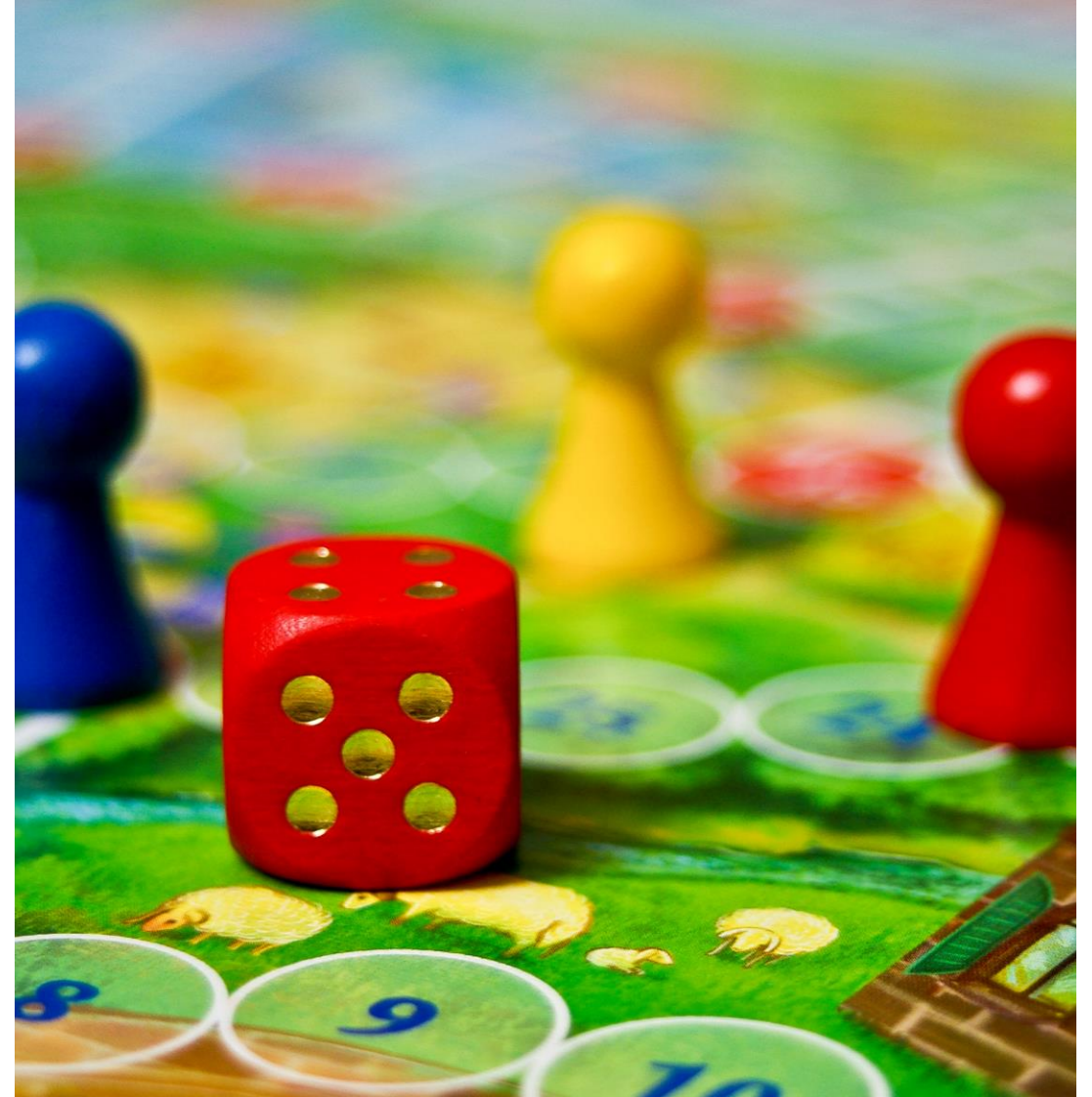
Re-Clustering the Initial Clusters

Same variables in re-clustering

***Worked with each initial cluster /
grouping separately***

*Avoids losing rare patterns amongst
common ones!*

([Salminen et al., 2021](#))



7 distinguishable RSE Personas generated from GH repository interactions data!

Low

- Ephemeral Contributor
- Occasional Contributor

Moderate

- Project Organiser
- Moderate Contributor

High

- Low-Process Closer
- Low-Coding Closer
- Active Contributor

Repository Contributions of RSE Personas

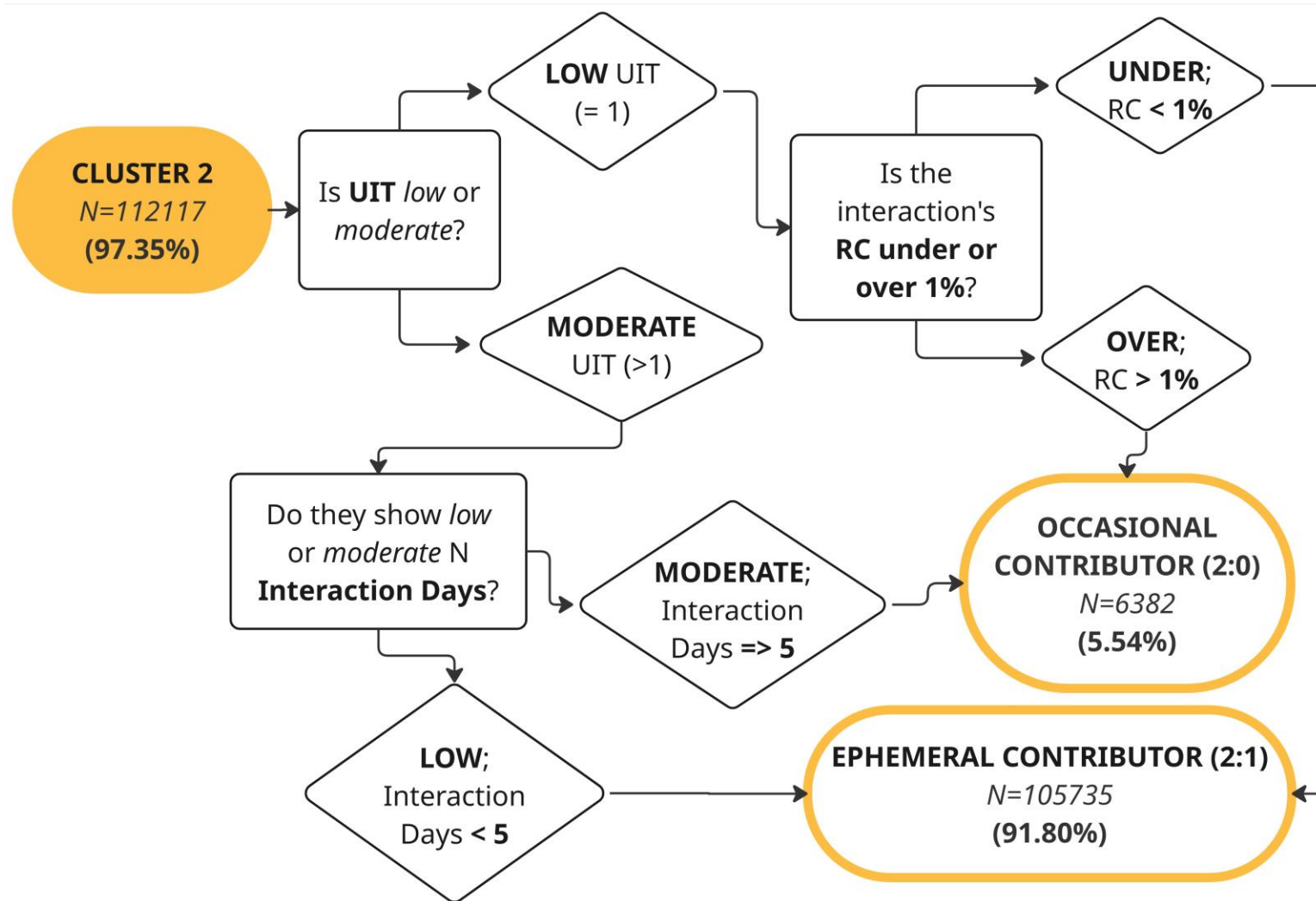
Moderate

High

Low

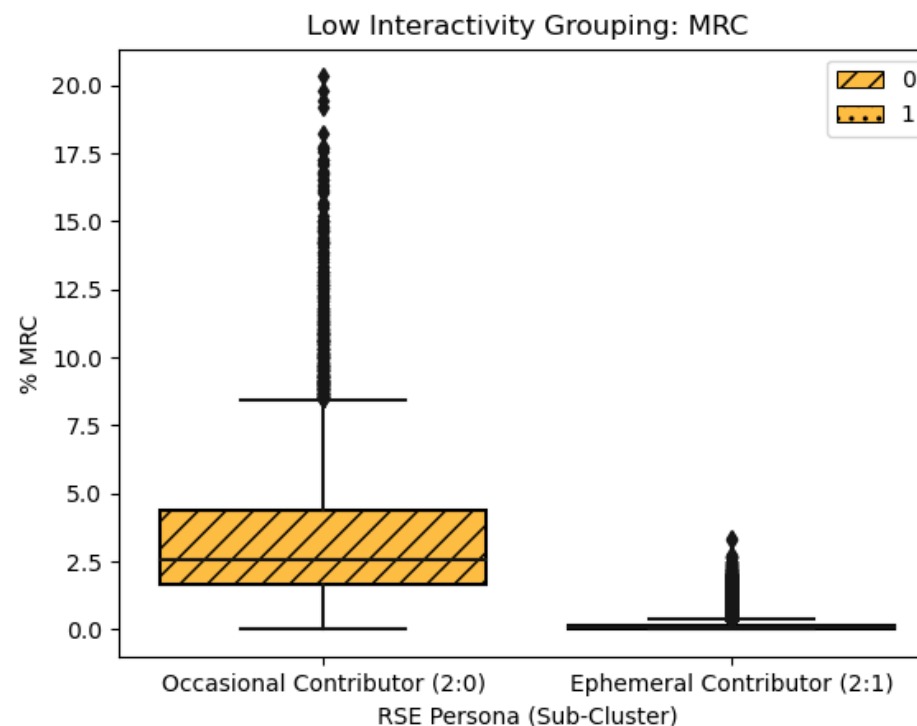
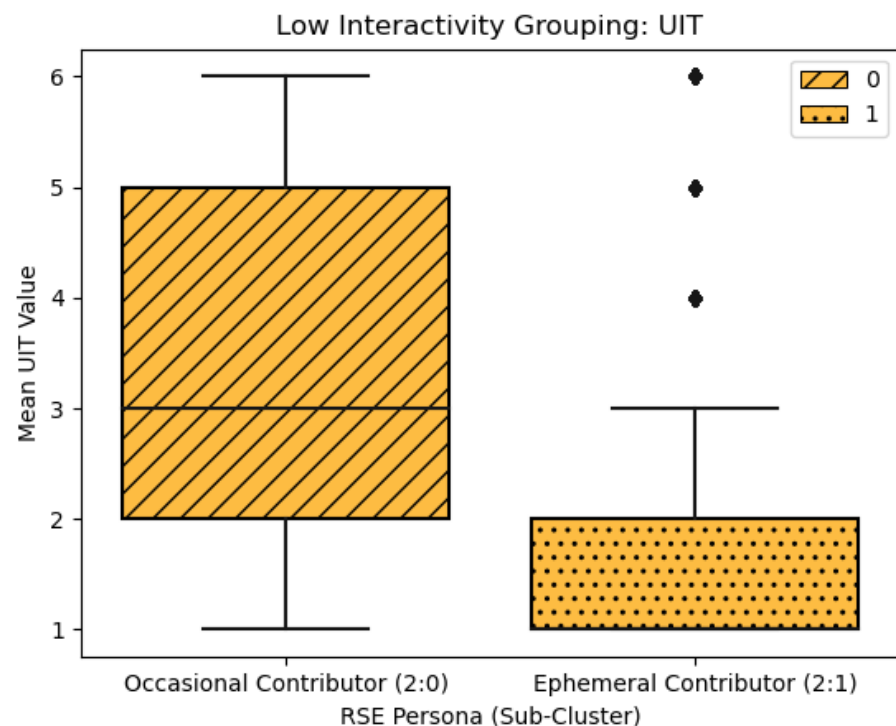
Variable \ Mean Values (%)	Sub-cluster 0:0	Sub-cluster 0:1	Sub-cluster 1:0	Sub-cluster 1:1	Sub-cluster 1:2	Sub-cluster 2:0	Sub-cluster 2:1
	Project Organiser	Moderate Contributor	Low-Coding Closer	Active Contributor	Low-Process Closer	Occasional Contributor	Ephemeral Contributor
RC Commit Creation	10.70	30.13	32.88	83.19	74.95	3.23	0.06
RC Issue Creation	12.35	22.47	40.47	51.72	16.12	3.61	0.38
RC Issue Closure	12.96	43.62	74.85	82.59	72.68	2.07	0.08
RC Assigned Issues	22.71	34.83	65.48	77.09	11.93	2.76	0.03
RC Pull Request Created	15.55	25.56	36.74	50.92	37.04	5.40	0.20
RC Pull Request Closed	16.11	44.81	71.64	86.24	84.13	1.94	0.03
MRC	14.85	31.32	50.08	69.10	42.54	3.41	0.15

Low Interactivity Personas



Low Interactivity Personas

Key Splits: UIT, MRC



Low Interactivity Personas

Ephemeral Contributor (91.80%)

Very Low Interactivity (2:1)

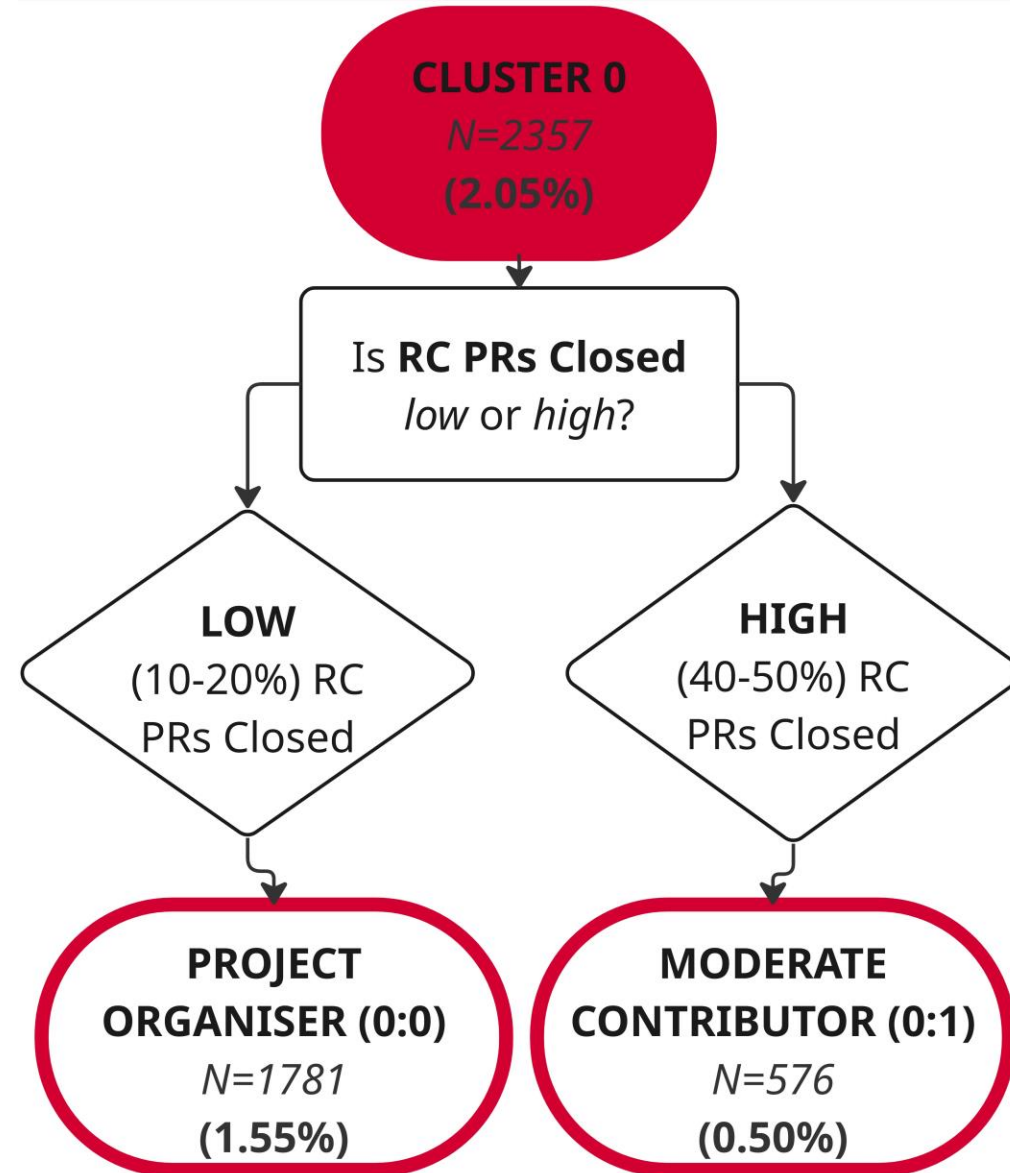
- Very low mean UIT (1.56) and MRC (0.15%)
- **Narrow and shallow contributions!**
- **Low net impact:** ~no net issues / PR interactions (-0.08%; 0.17%)
- **"ephemeral"**: only 3 mean interaction days and 0.23% RC, across an interaction period of 124 days (~4 months)
- May visit only to **request a fix or new feature** (higher issue creation than commits or closure of issues), but plenty of them: contribute lots of ideas

Occasional Contributor (5.54%)

Low Interactivity (2:0)

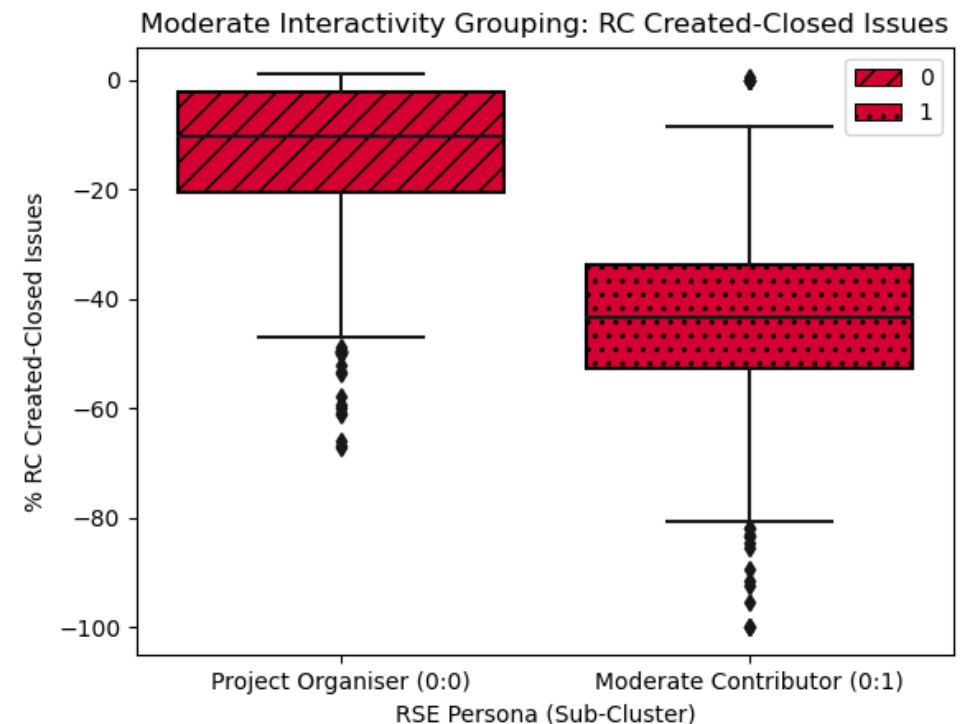
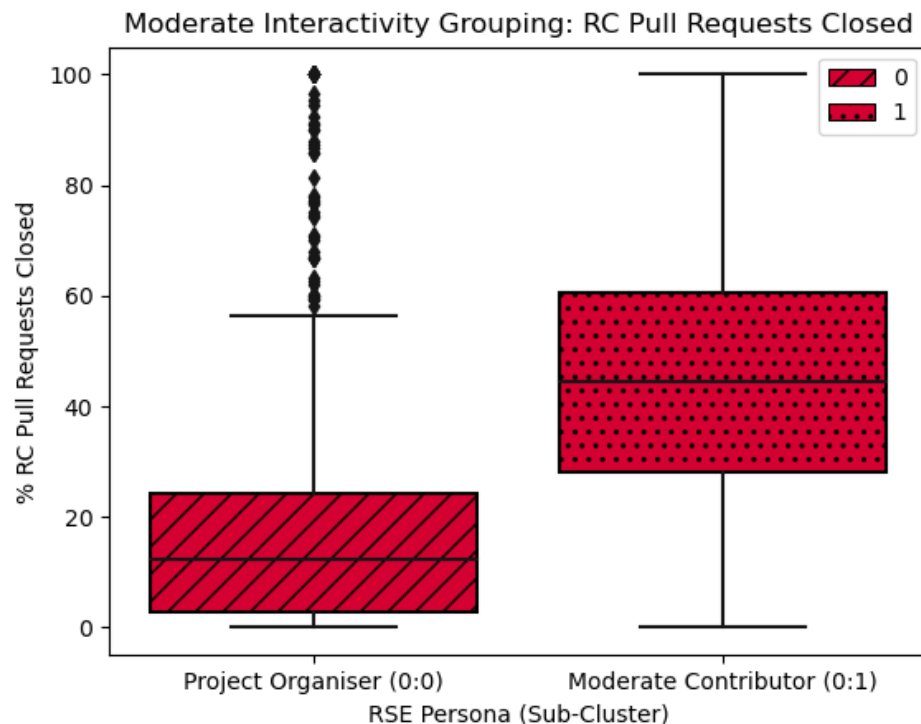
- UIT mean is 3.42%, Mean MRC 3.41%
- **low general contributions across moderate variety...**
- Weak **net** RC Issue **Closure** (-2.03%) but weak **net Creation** of PRs (3.45%)
- **~4% of all interaction days** in their repo (mean 43), therefore **"occasional contributor"** across interaction period of 802.65 days (2.20 years)
- **Higher frequency than rarer but higher-interaction personas**, so still important for significant RS development within their projects

Moderate Interactivity Personas



Moderate Interactivity Personas

Key Splits: PRs Closed, (Net) Created-Closed Issues RC



Moderate Interactivity Personas

Project Organiser (1.55%)

Low-Moderate Interactivity (0:0)

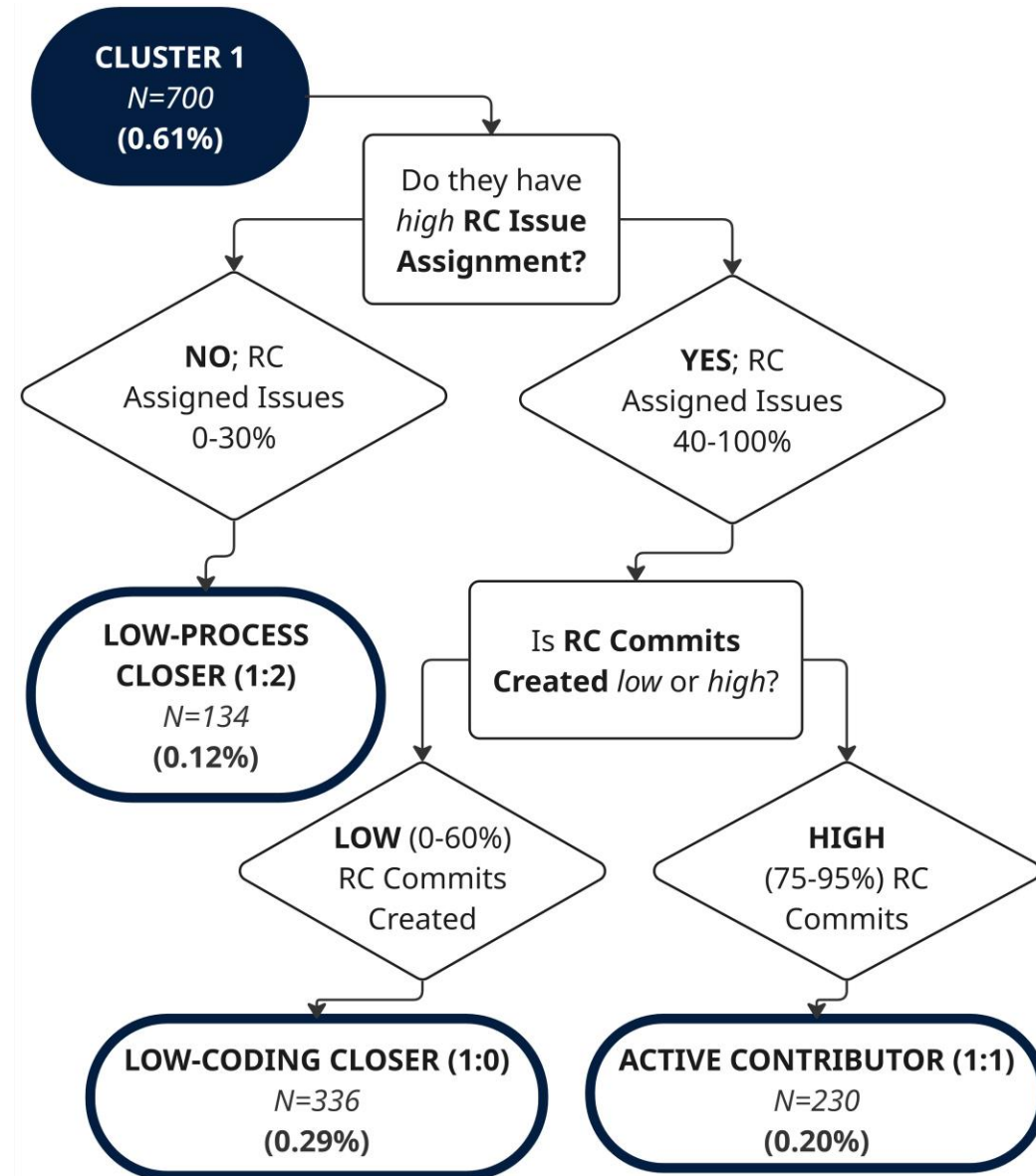
- Moderately high mean UIT: 4.88; MRC still low: 14.85%
- **Varied (wide) but very shallow** engagement with the repo...
- Only 10.70% RC for Commit Creation, but 22.71% of Assignment to Issue Tickets (range: 12.01%)
- **High assignment** but relatively low Issue/PR closure rates may mean "**keep me in the loop**"?
- **Moderate time involvement** (119.97 interaction days across 3.85 years)
- **Matches 'low MRC, high UIT - Project Manager' role** initially expected at outset of pilot study (renamed)
- Focus on **managing projects and development effort** instead of active development role?

Moderate Contributor (0.50%)

Moderate Interactivity (0:1)

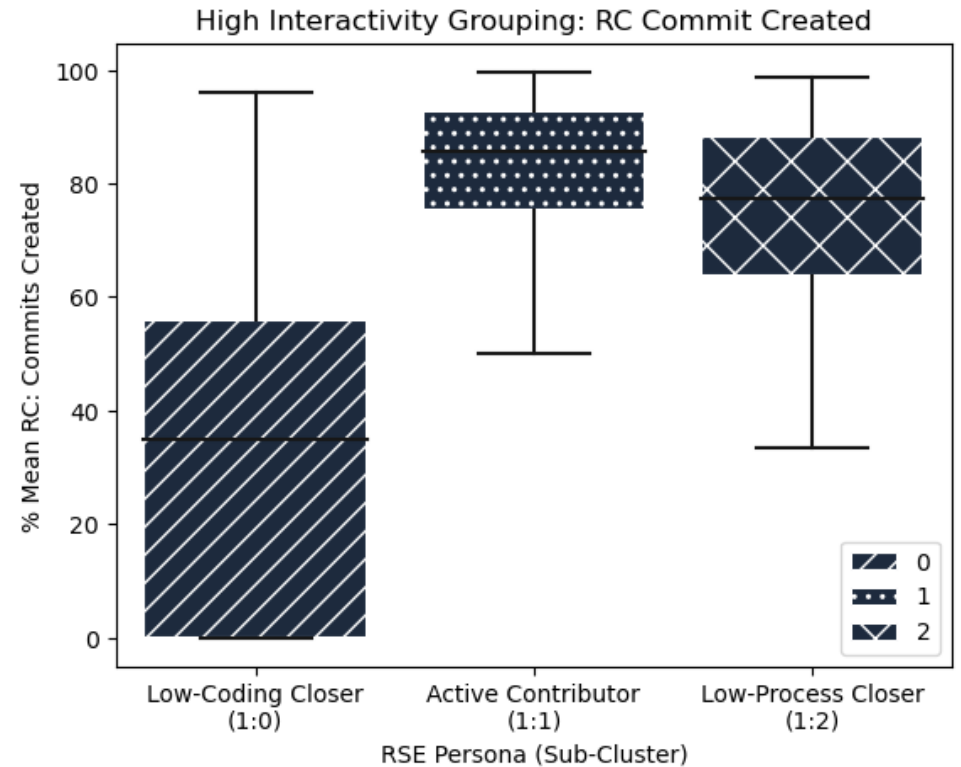
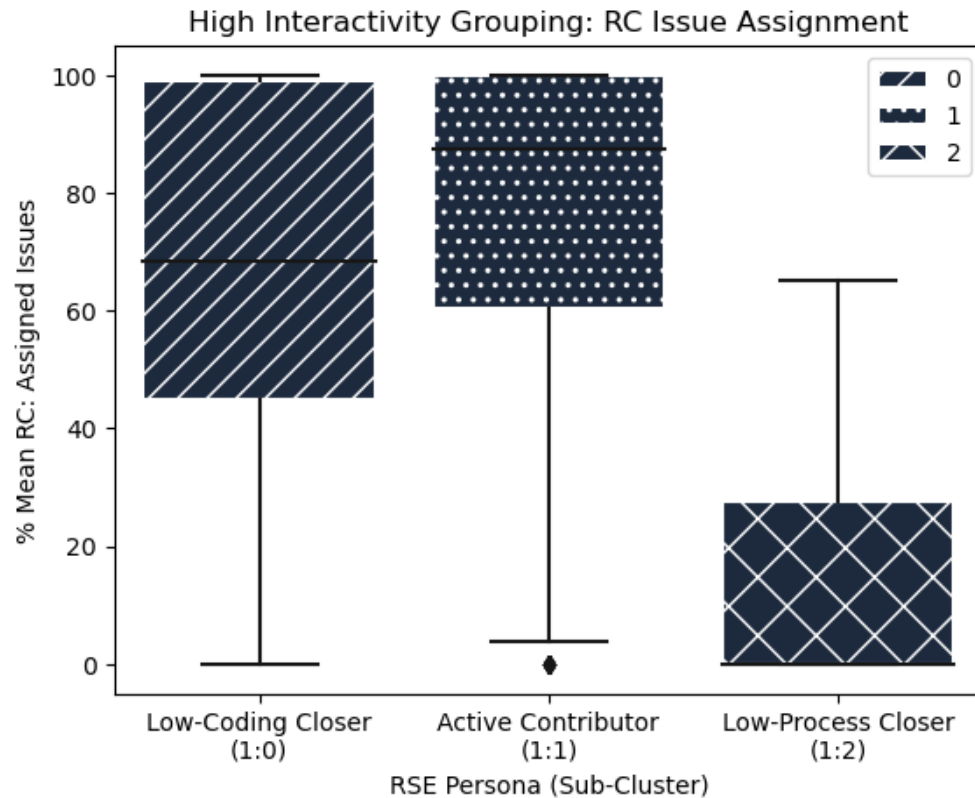
- High mean UIT (5.45) and moderate MRC: 31.32%
- **Varied but not too deep** contributions generally
- RC ranges moderate too: 22.34%
- **Focussed on net Closure!** RC values for PR and Issue **Closure nearly double equivalent Creation** types, leading to net Issue Closure of -43.40%; net closure of PRs of -19.25%
- 278 Interaction days, across 2140 days (5.86 years) Interaction Period
- **Uses dev management features frequently, but also does the dev work required to close items:** 30.13% RC Commit Creation (c.f. Project Organisers!)

High Interactivity Personas



High Interactivity Personas

Key Splits: RC Issue Assignment, RC Commits Created



High Interactivity Personas

Low-Process Closer (0.12%)

Moderate-High Interactivity (1:2)

- High UIT 5.15; moderate MRC 42.54%, but...
- Strong **behaviour preference away from Assignment/Issue Creation** (11.93% and 16.12%) **towards PR Closure** (84.13%)!
- **Highest net PR closure** (37.04% created minus closed 84.13% = net -47.10%) and **strong net ticket closure** (16.12% created minus closed 72.68% = net -56.56%) **due to low opening rates**
- **Showing up!** 347.72 mean interaction days; Interaction Period 2573 days (7.05 years); 53.5% of repositories' total
- Likely a fixer, **keener on 'getting things done' by closing existing tickets/PRs** than opening new ones ("**low-process**")

Low-Coding Closer (0.29%)

High Interactivity (1:0)

- UIT 5.43; MRC 50.08%
- **High variety of interactions, good volume!**
- More **consistent RCs** than Low-Process Closers (range 41.97% c.f. 72.20%!)
- **Low commit creation** RC: 32.88%, (therefore "**low-coding**") ...but...
- Still **high closure rates?! 74.85% Issue Closure** and **71.64% PR Closure**
- **Present!** 286.35 interaction days on average; Interaction Period 2362 days (6.47 years); **42.47%** of repo's total days
- May be **triaging PRs and Issue Tickets**, closing duplicated/irrelevant items or working on items needing no commit creation to resolve them?

High Interactivity

Active Contributor (0.20%)

Very High Interactivity (1:1)

- Highest mean UIT (5.88) and highest MRC (69.10%)
- **High variety and deep volume of contributions!**
- Highest **Issue Ticket Assignment** of all personas (77.09% of assignments in their repos to these repo-individuals)
- Great **net closure of PRs**: -82.07% (RC Created Minus Closed Issues)
- Over **65% of all Interaction Days** in their repo by them (397 Interaction Days) across Interaction Period of nearly 7 years (2523.62 days)
- High usage of **development management features** (issue tickets, PRs, assignment) AND **impressive codebase contributions** through commits
- Important **core members** of their repos
- Matches hypothesised persona!

Limitations

Variable Selection

- UIT too simplistic, **MRC is ok summary** (with caveats)?
- **"High Responsibility" Interaction Types** (such as Assignment or PR Closure) important
- Commit Classification Methods ([Vasilescu et al., 2014](#) or [Hattori-Lanza, 2008](#)) **not different** (commit size, file type, or message key words)

RS Repos vs Projects

- **Forks discounted** for on collaborative coding, but [Kalliamvakou et al., 2016](#) include all forks, working at 'project' level
- 'Offline' work and external tools...

Skewed towards Best Practices?

- Zenodo research repository – **repos polished before publishing?**

Limitations

Bots and RSE Personas

- Bots not excluded after pilot study

Missing RSE Personas?

- **"Focused Developer"** type hypothesised but not found (low UIT and high MRC percentages)
- Artifact of the MRC methodology?

Not All Interactions Are Equal:

- Not applying weighting for rarity, importance or specific types
 - But... % RC (**Relative Contribution**) IS comparable measure between repo-individuals within a specific repository
-

Limitations

Quantity Is Not The Same As Quality

- **Not assessing quality of interactions**, only quantity
- Cannot estimate **effectiveness** of RSE behaviour patterns without further work...

Hidden Contributors:

- **Missing non-coding roles** such as project managers or researchers...

RSE Persona Dynamics:

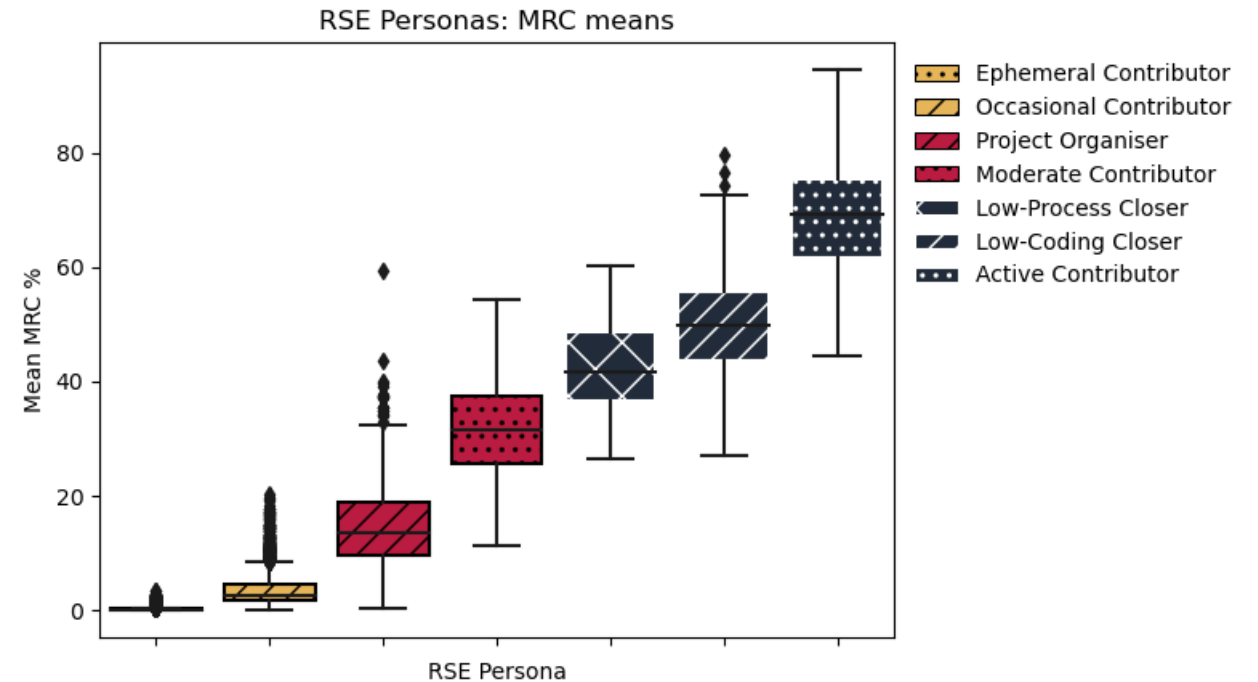
- Not all Personas are found in same repos: more work required to explore **distributions and dynamics**
-

Conclusions

Novel method, successfully applied to nearly **4 million interactions** from **115,174 repo-individuals** across 1,284 RS repos on GH

Seven identified RSE Personas describe common and rare contribution patterns in RSE

Persona relationships can be explained in terms of **low, moderate and high interactivity** groupings

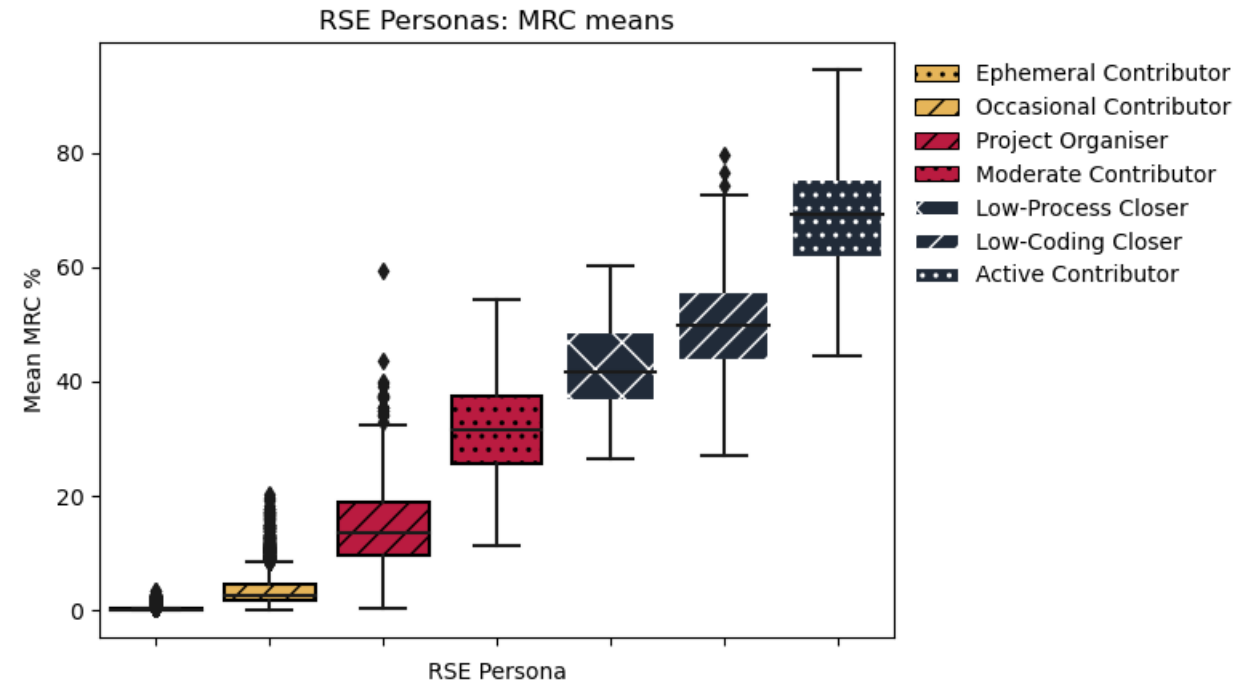


Conclusions

Hypothesised **Active Contributor**, **Occasional Contributor** and **Project Organiser RSE** Personas confirmed!

Four additional RSE Personas identified:

- **Ephemeral Contributor**
- **Moderate Contributor**
- **Low-Process Closer**
- and **Low-Coding Closer**



Next Steps... (this research)



- **Link with other research** such as RS project categorisation, or mixed methods studies?
RSE Persona Dynamics!
- **Planning a tool** for analysing RSE Personas for specific GH repos or finding your own RSE Persona
- **Extended Abstract** for SERS26 workshop: *using ML for RSE Persona prediction!*

Preprint: doi.org/10.48550/arXiv.2510.05390

This work was supported by a
Doctoral Training Partnership award
for project number 266270 through
grant EP/T517884/1

With thanks to EPCC colleagues,
RDRC and deRSE25 attendees.



github.com/FlicAnderson/RSE-personas

