Unveiling Experts in Data Science for Team Composition: Insights from Mining Software Repositories

José Antônio
Eduardo Figueiredo
Johnatan Oliveira

LabSoft Seminar. Jun 26th, 2025

Characterize Experts in Data Science

- Explore the activity of software repositories
  - Identify experts in data science and programming, that we call data science programmers.
- Identify their roles in team composition

Research Questions

- RQ1: What information about data science programmers do commit-based metrics provide?

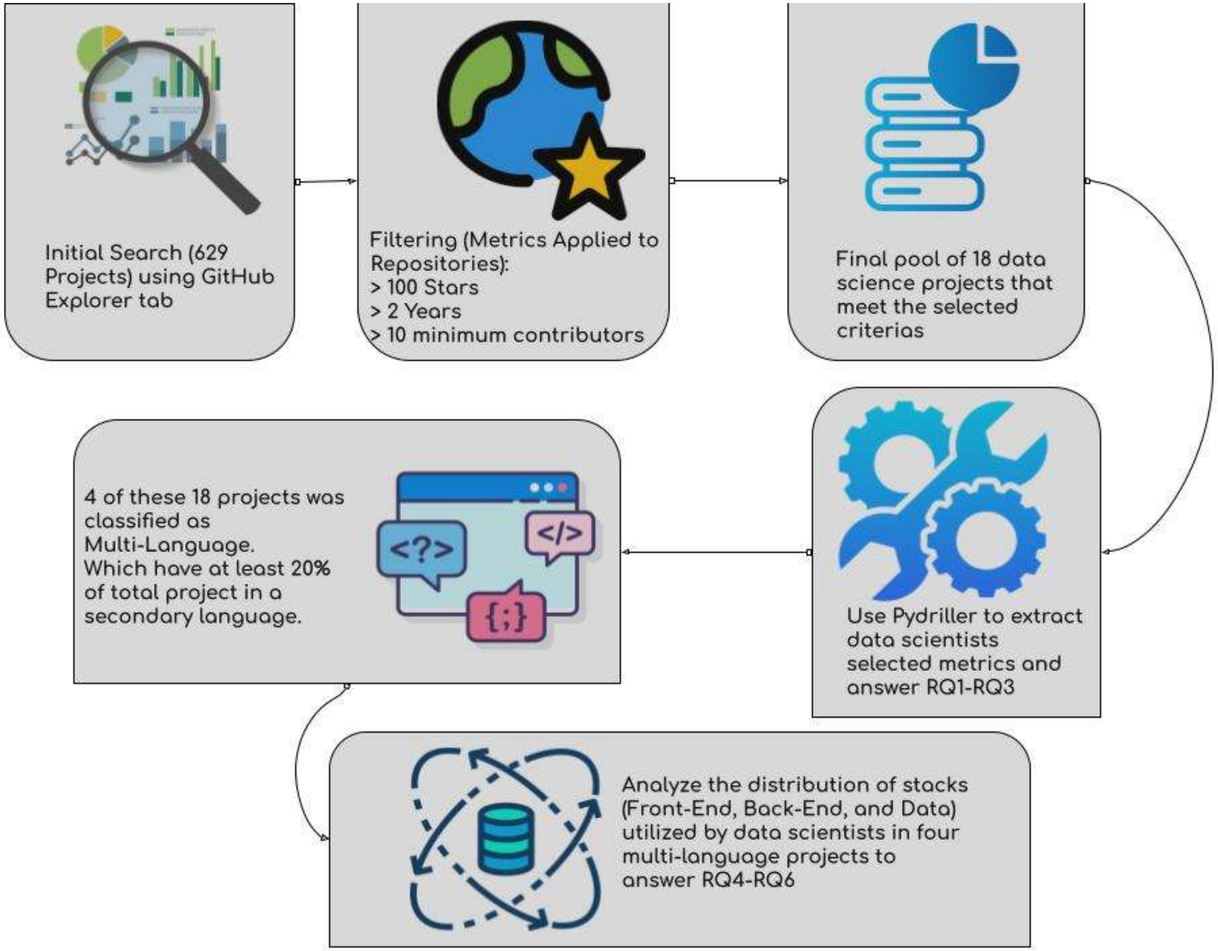
- RQ2: What are the characteristics of experts in data science software projects?

- RQ3: Which programming languages are mostly used by data science programmers?

Research Questions

- RQ4: What is the general composition of project teams?
- RQ5: What are the roles and responsibilities of experts within the project teams?
- RQ6: What are the defining characteristics of the project teams?

Steps for Criteria Evaluation



Initial Search (629 Projects) using GitHub Explorer tab

Filtering (Metrics Applied to Repositories):
> 100 Stars
> 2 Years
> 10 minimum contributors

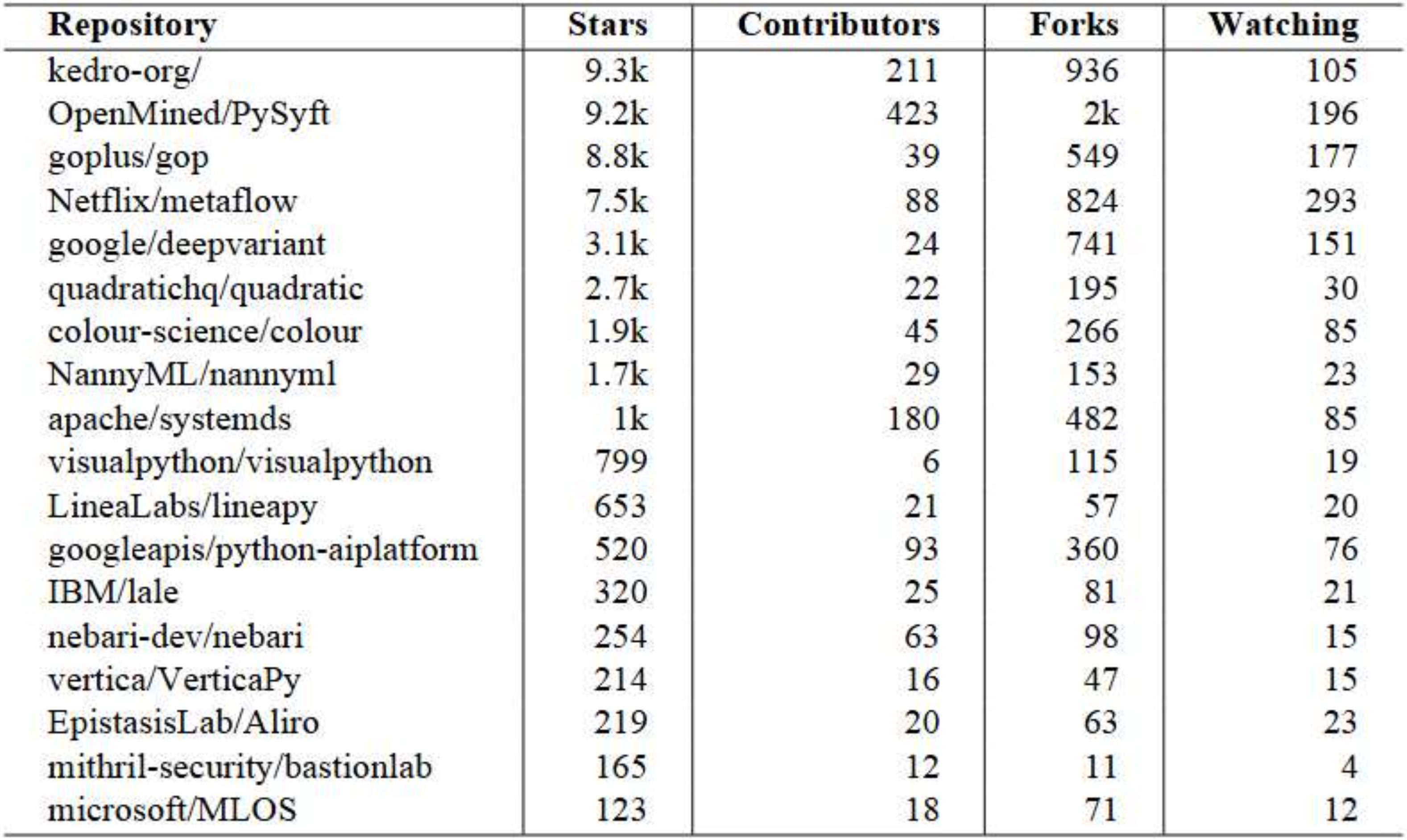
Final pool of 18 data science projects that meet the selected criterias

Use Pydriller to extract data scientists selected metrics and answer RQ1-RQ3

4 of these 18 projects was classified as Multi-Language. Which have at least 20% of total project in a secondary language.

Analyze the distribution of stacks (Front-End, Back-End, and Data) utilized by data scientists in four multi-language projects to answer RQ4-RQ6

Used GitHub Repositories



Repository	Stars	Contributors	Forks	Watching
kedro-org/	9.3k	211	936	105
OpenMined/PySyft	9.2k	423	2k	196
goplus/gop	8.8k	39	549	177
Netflix/metaflow	7.5k	88	824	293
google/deepvariant	3.1k	24	741	151
quadratichq/quadratic	2.7k	22	195	30
colour-science/colour	1.9k	45	266	85
NannyML/nannyml	1.7k	29	153	23
apache/systemds	1k	180	482	85
visualpython/visualpython	799	6	115	19
LineaLabs/lineapy	653	21	57	20
googleapis/python-aiplatform	520	93	360	76
IBM/lale	320	25	81	21
nebari-dev/nebari	254	63	98	15
vertica/VerticaPy	214	16	47	15
EpistasisLab/Aliro	219	20	63	23
mithril-security/bastionlab	165	12	11	4
microsoft/MLOS	123	18	71	12

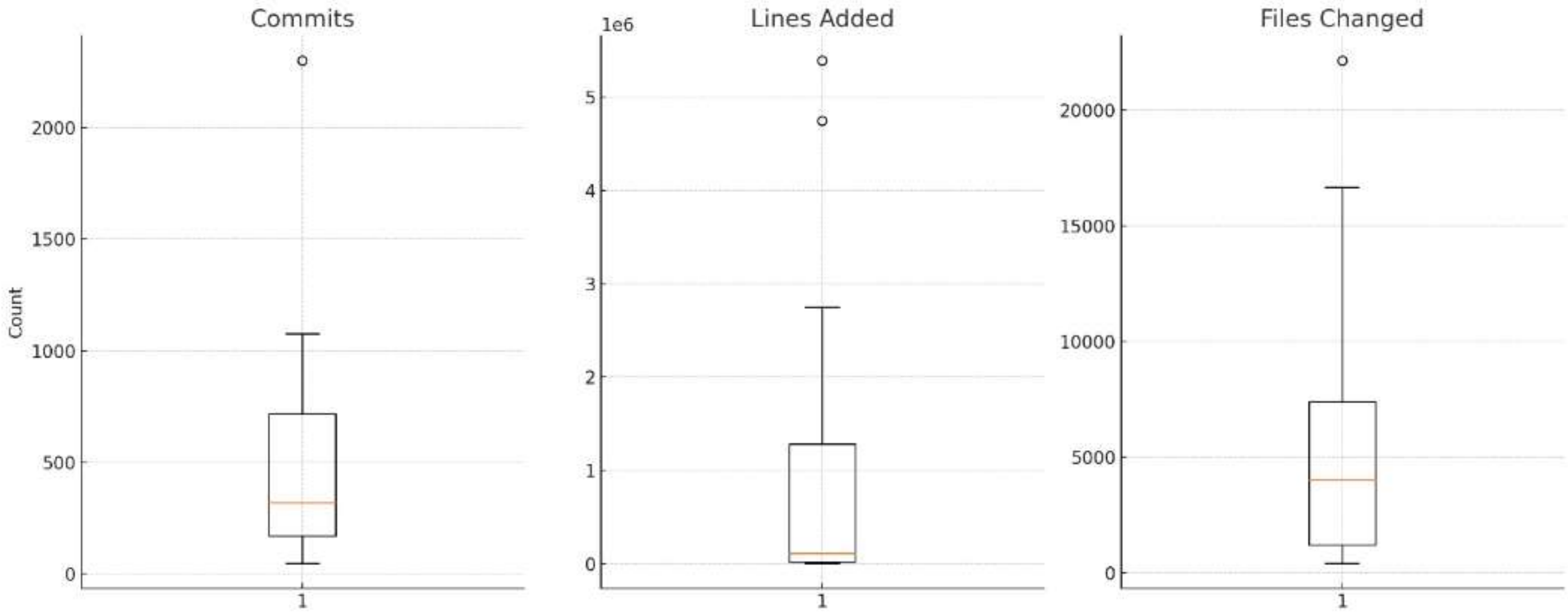
Metrics

- Collect data with Pydriller, including developer emails and commit messages.

- Number of Commits
- Lines of Code
- Number of Changed Files

Results RQ1

- RQ1: What information about data science programmers do commit-based metrics provide?



Results RQ2

- RQ2: What are the characteristics of individuals identified as experts in data science projects?
- Analysis of 69 Data Scientists across 18 projects

- Progressive Data Scientist (53% of all DS)
  - Spend 60% in additions and no more than 40% in deletions
- Ordinary Data Scientist (25% of all DS)
  - Not exceed 60% in additions or deletions of their total activity
- Conservative Data Scientist (22% of all DS)
  - Spend 60% in deletions and no more than 40% in additions

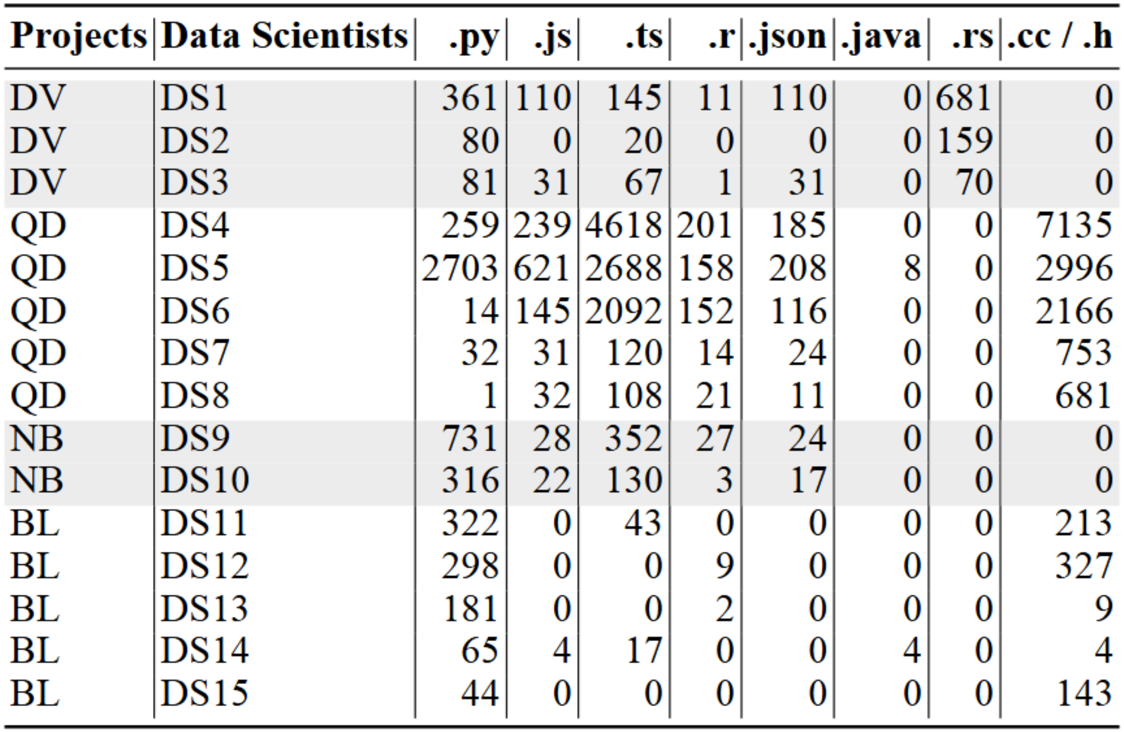
Results RQ2

- RQ2: What are the characteristics of individuals identified as experts in data science projects?

- Mono-Language Data Scientists (68% of all DS)
  - No more than 40% of their work performed on a secondary language
- Multi-Language Data Scientists (32% of all DS)
  - At least 40% of their work performed on a secondary language

Results RQ4

- RQ4: What is the general composition of project teams?



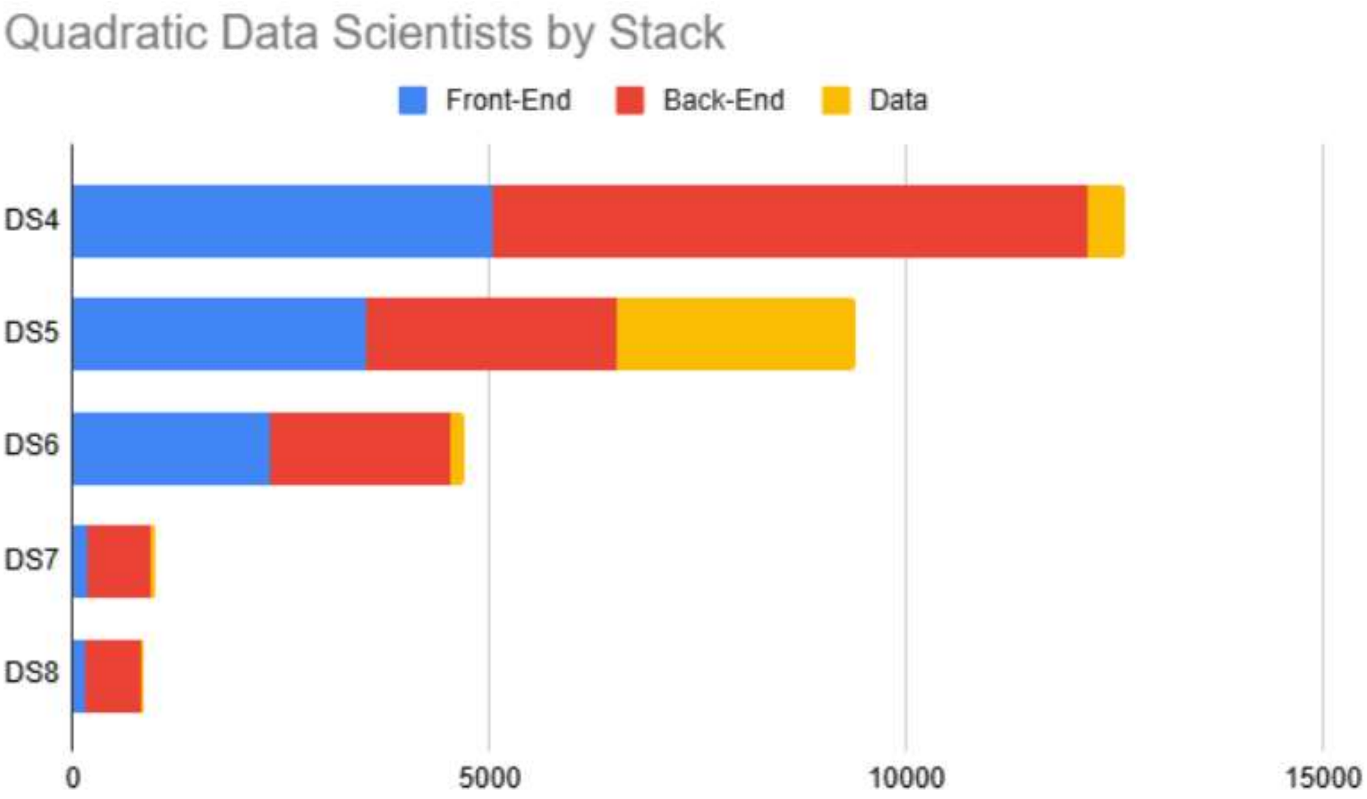
Projects	Data Scientists	.py	.js	.ts	.r	.json	.java	.rs	.cc / .h
DV	DS1	361	110	145	11	110	0	681	0
DV	DS2	80	0	20	0	0	0	159	0
DV	DS3	81	31	67	1	31	0	70	0
QD	DS4	259	239	4618	201	185	0	0	7135
QD	DS5	2703	621	2688	158	208	8	0	2996
QD	DS6	14	145	2092	152	116	0	0	2166
QD	DS7	32	31	120	14	24	0	0	753
QD	DS8	1	32	108	21	11	0	0	681
NB	DS9	731	28	352	27	24	0	0	0
NB	DS10	316	22	130	3	17	0	0	0
BL	DS11	322	0	43	0	0	0	0	213
BL	DS12	298	0	0	9	0	0	0	327
BL	DS13	181	0	0	2	0	0	0	9
BL	DS14	65	4	17	0	0	4	0	4
BL	DS15	44	0	0	0	0	0	0	143

Results RQ5

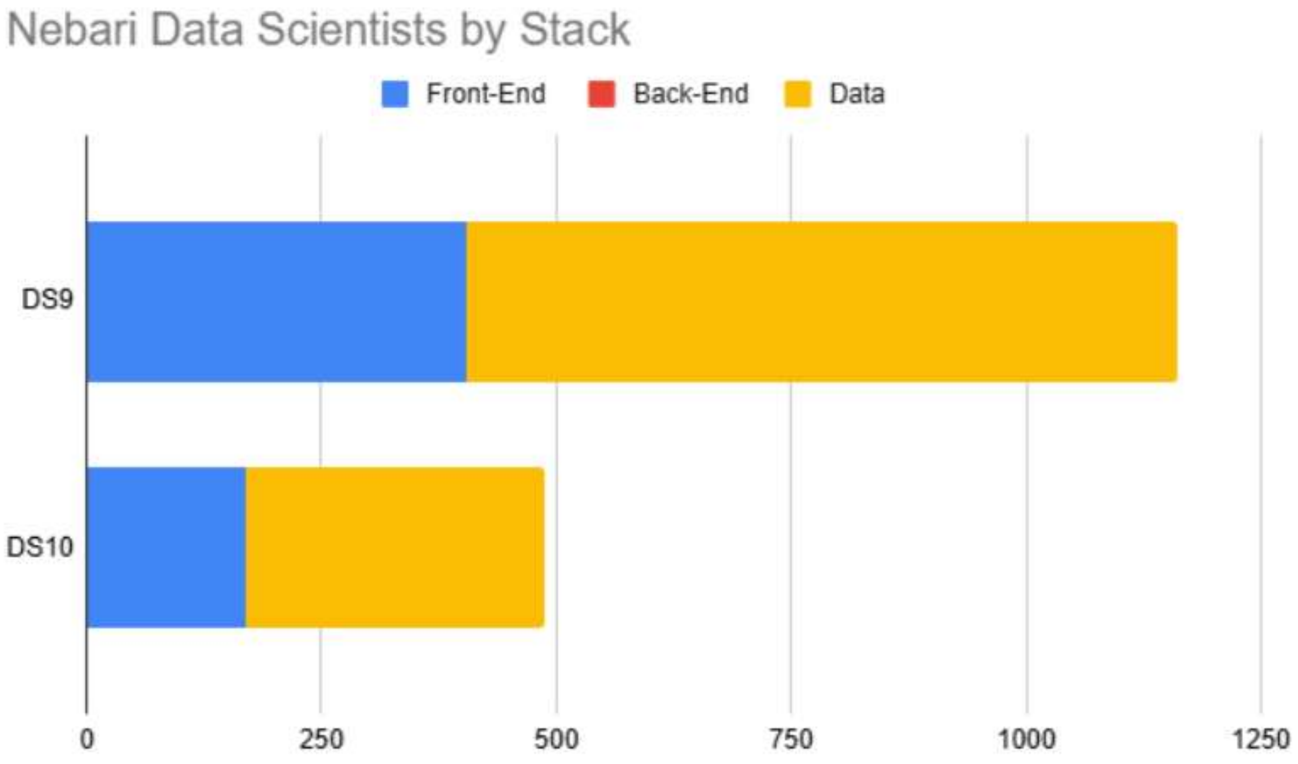
- RQ5: What are the DS's roles and responsibilities within the project teams?

- (Front-End, Back-End, and Data)
  - Multi-language projects: Quadratic, Nebari, BastionLab and DeepVariant

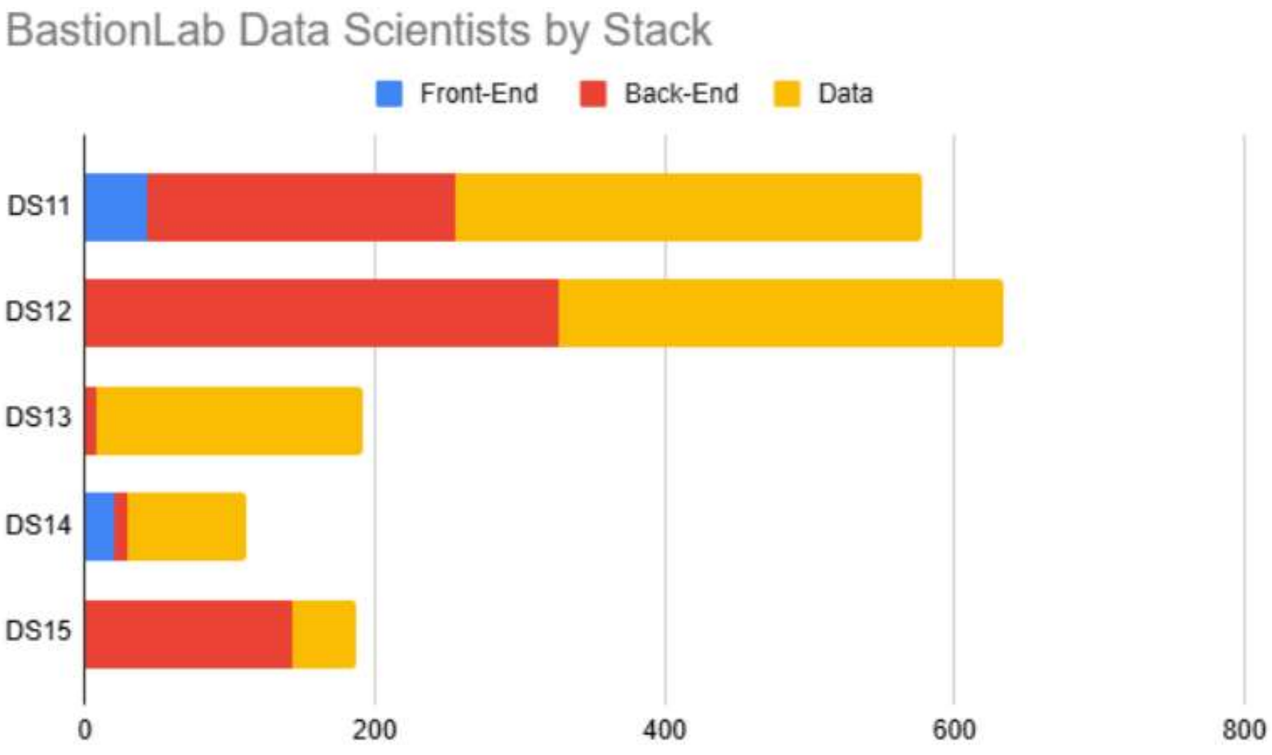
Results RQ5


Quadratic Data Scientists by Stack
Front-End
Back-End
Data
DS4
DS5
DS6
DS7
DS8
0
5000
10000
15000

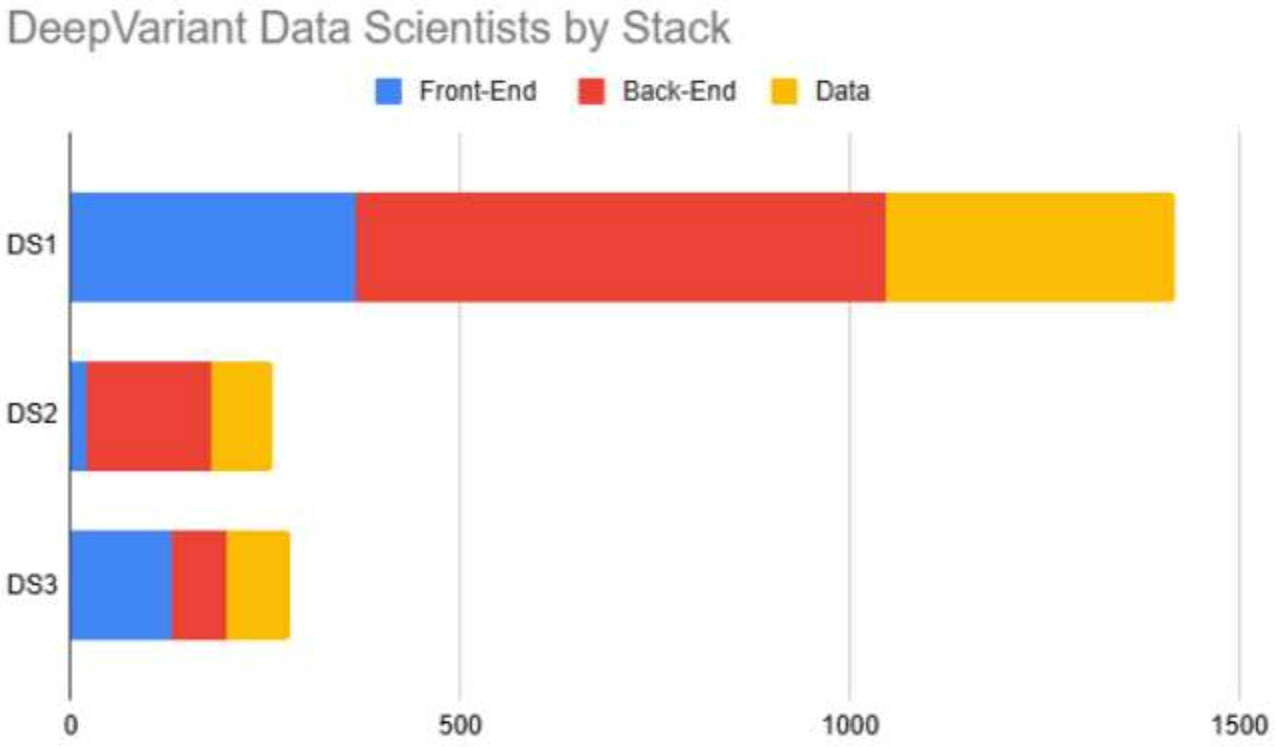
Results RQ5


Nebari Data Scientists by Stack
🟦 Front-End
🟥 Back-End
🟨 Data
DS9
DS10
0
250
500
750
1000
1250

Results RQ5


BastionLab Data Scientists by Stack
Front-End
Back-End
Data
DS11
DS12
DS13
DS14
DS15
0
200
400
600
800

Results RQ5


DeepVariant Data Scientists by Stack
Front-End
Back-End
Data
DS1
DS2
DS3
0
500
1000
1500

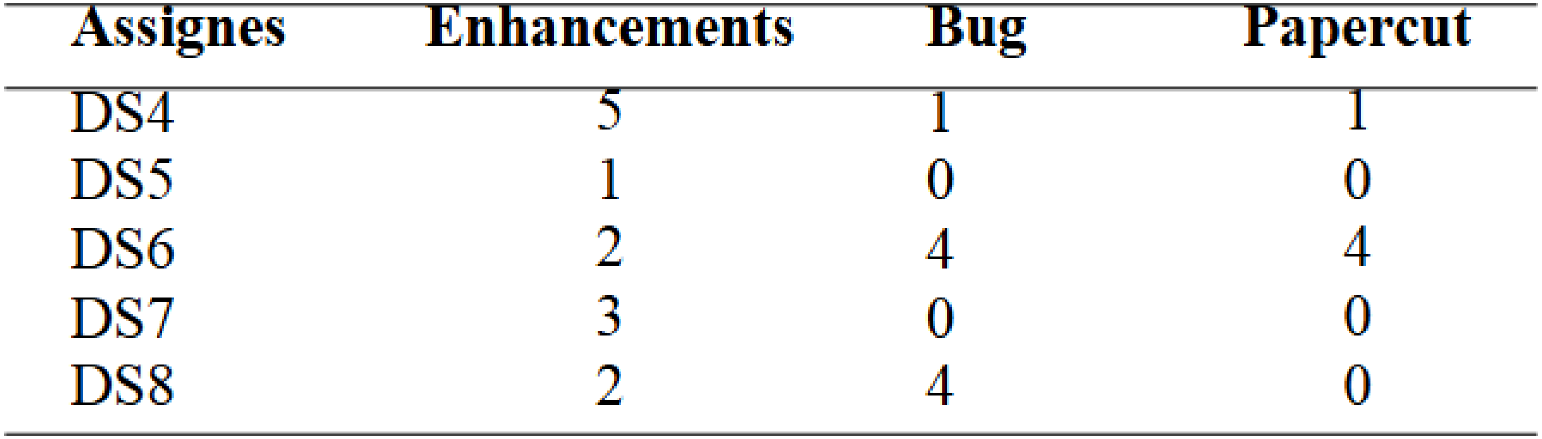
Results RQ6

- Closer look at Quadratic

- We selected Quadratic because it has a clear workflow of tasks assigned to developers.

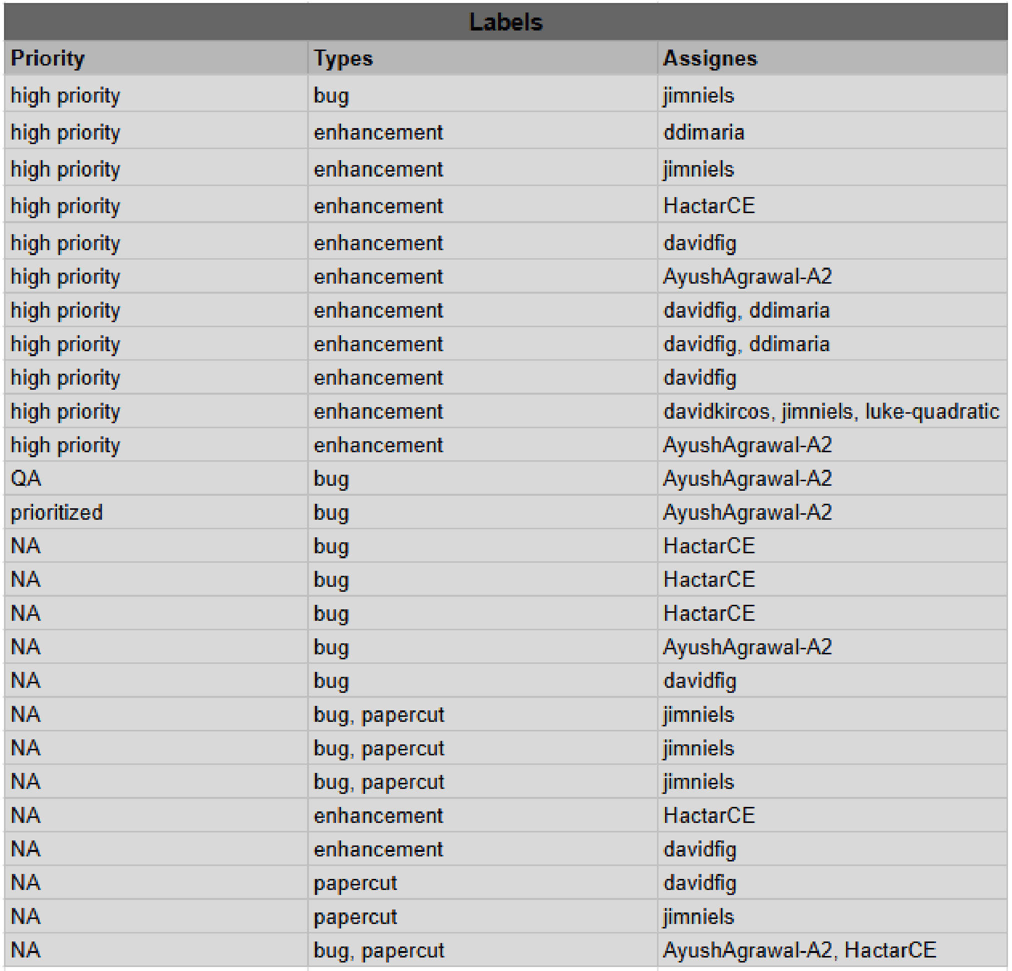
Results RQ6

- Enhancements - New Feature or Request
- Bug - something is broken
- Papercut - annoying but not the end of the world



Assignes	Enhancements	Bug	Papercut
DS4	5	1	1
DS5	1	0	0
DS6	2	4	4
DS7	3	0	0
DS8	2	4	0

Results RQ6



Labels		
Priority	Types	Assignes
high priority	bug	jimniels
high priority	enhancement	ddimaria
high priority	enhancement	jimniels
high priority	enhancement	HactarCE
high priority	enhancement	davidfig
high priority	enhancement	AyushAgrawal-A2
high priority	enhancement	davidfig, ddimaria
high priority	enhancement	davidfig, ddimaria
high priority	enhancement	davidfig
high priority	enhancement	davidkircos, jimniels, luke-quadratic
high priority	enhancement	AyushAgrawal-A2
QA	bug	AyushAgrawal-A2
prioritized	bug	AyushAgrawal-A2
NA	bug	HactarCE
NA	bug	HactarCE
NA	bug	HactarCE
NA	bug	AyushAgrawal-A2
NA	bug	davidfig
NA	bug, papercut	jimniels
NA	bug, papercut	jimniels
NA	bug, papercut	jimniels
NA	enhancement	HactarCE
NA	enhancement	davidfig
NA	papercut	davidfig
NA	papercut	jimniels
NA	bug, papercut	AyushAgrawal-A2, HactarCE

Threats to Validity

- Small sample of Projects
- Developers with fewer commits but impactful contributions maybe undervalued
- Number of Stars and Contributors

Conclusion

- Future work could address these limitations by expanding the dataset.
- Incorporating more diverse metrics.

- How LLM would classify data scientists

Obrigado!