Meta-analysis of the machine learning operations open source ecosystem

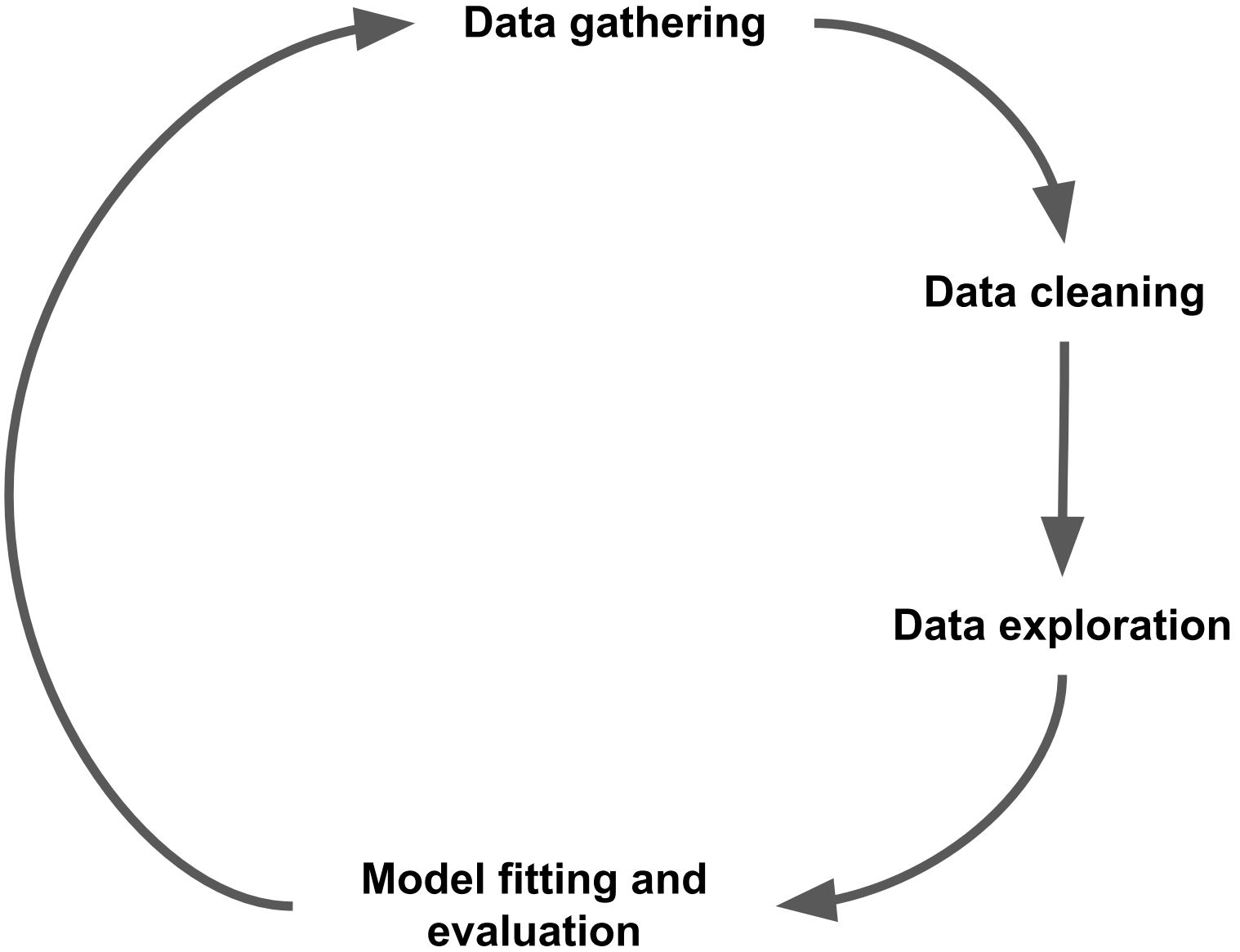
Florida Polytechnic University

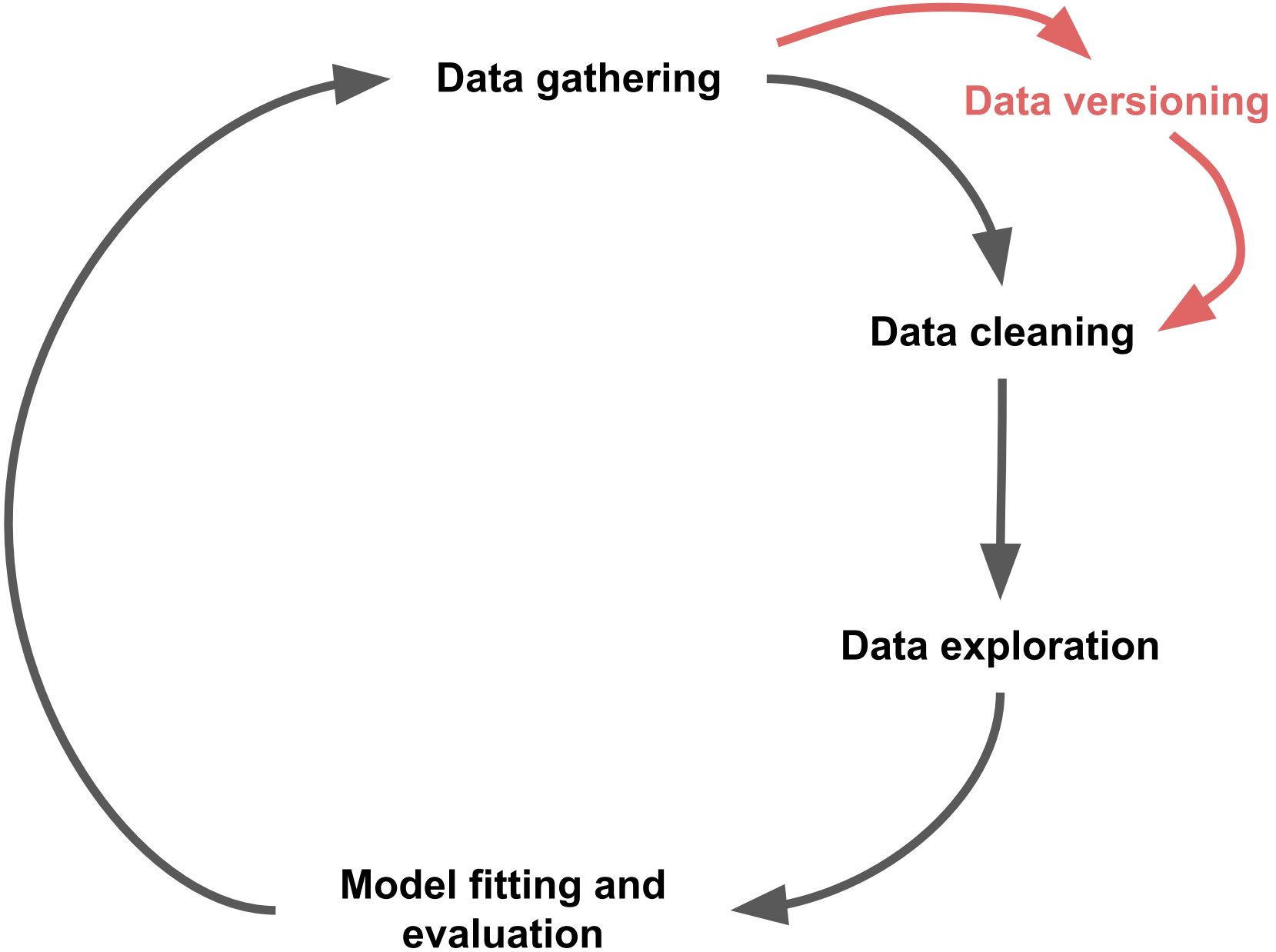
Isabel Zimmerman

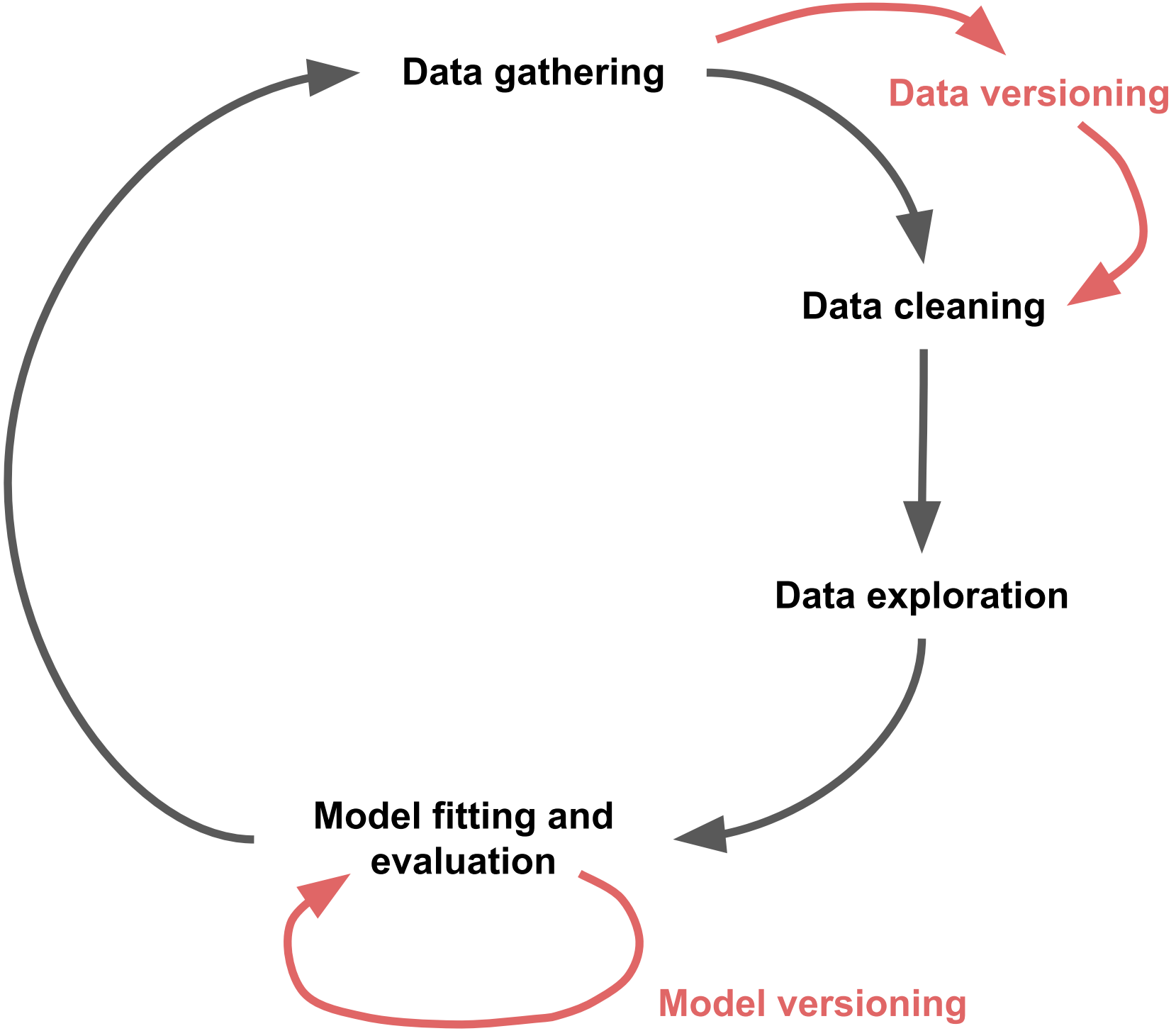
2023-04-26

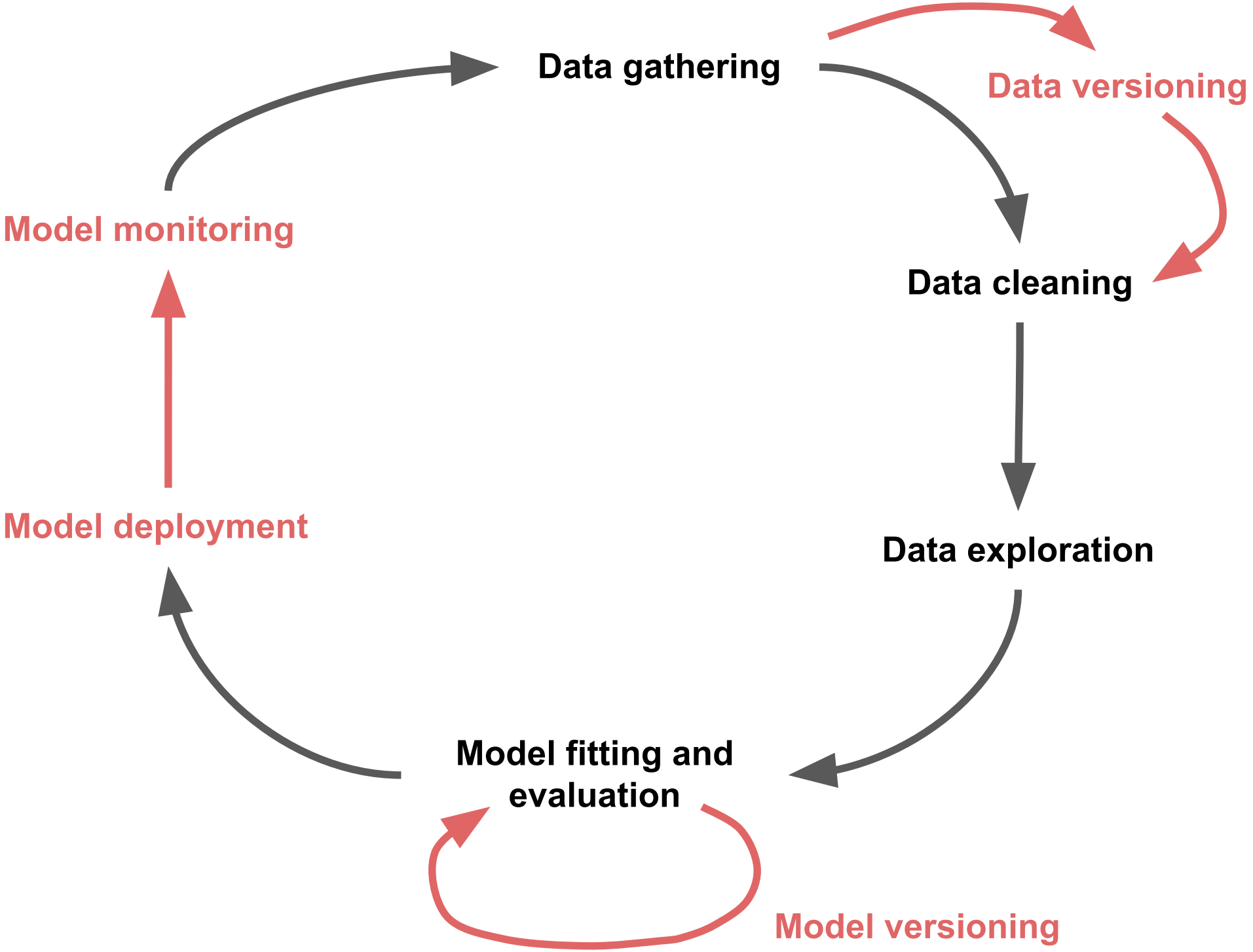
What is MLOps?

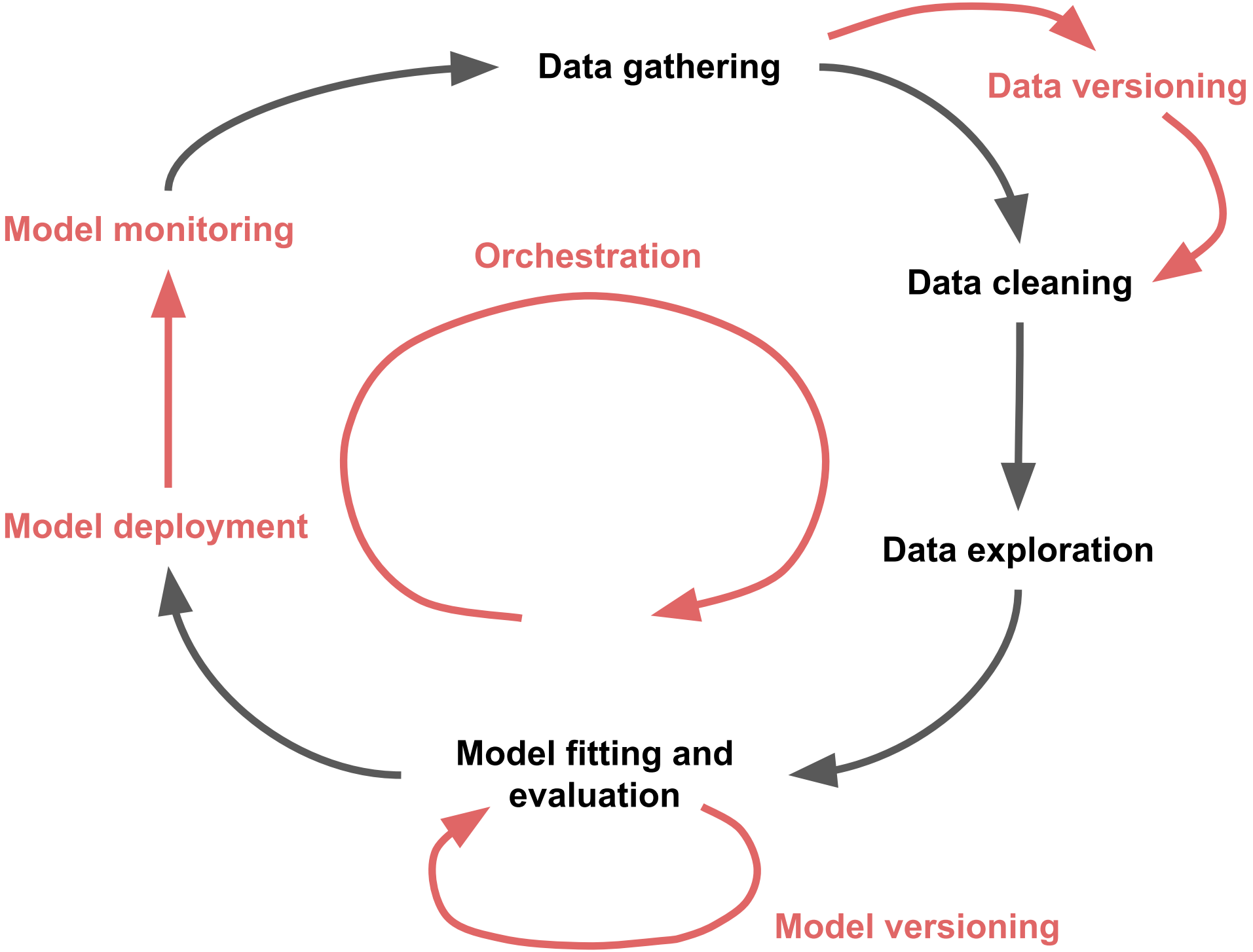
- set of practices to deploy and maintain machine learning models in production reliably and efficiently
- roots in Development Operations


Data gathering
Data cleaning
Data exploration
Model fitting and evaluation


Data gathering
Data versioning
Data cleaning
Data exploration
Model fitting and evaluation


Data gathering
Data versioning
Data cleaning
Data exploration
Model fitting and evaluation
Model versioning


Data gathering
Data versioning
Data cleaning
Data exploration
Model fitting and evaluation
Model versioning
Model deployment
Model monitoring

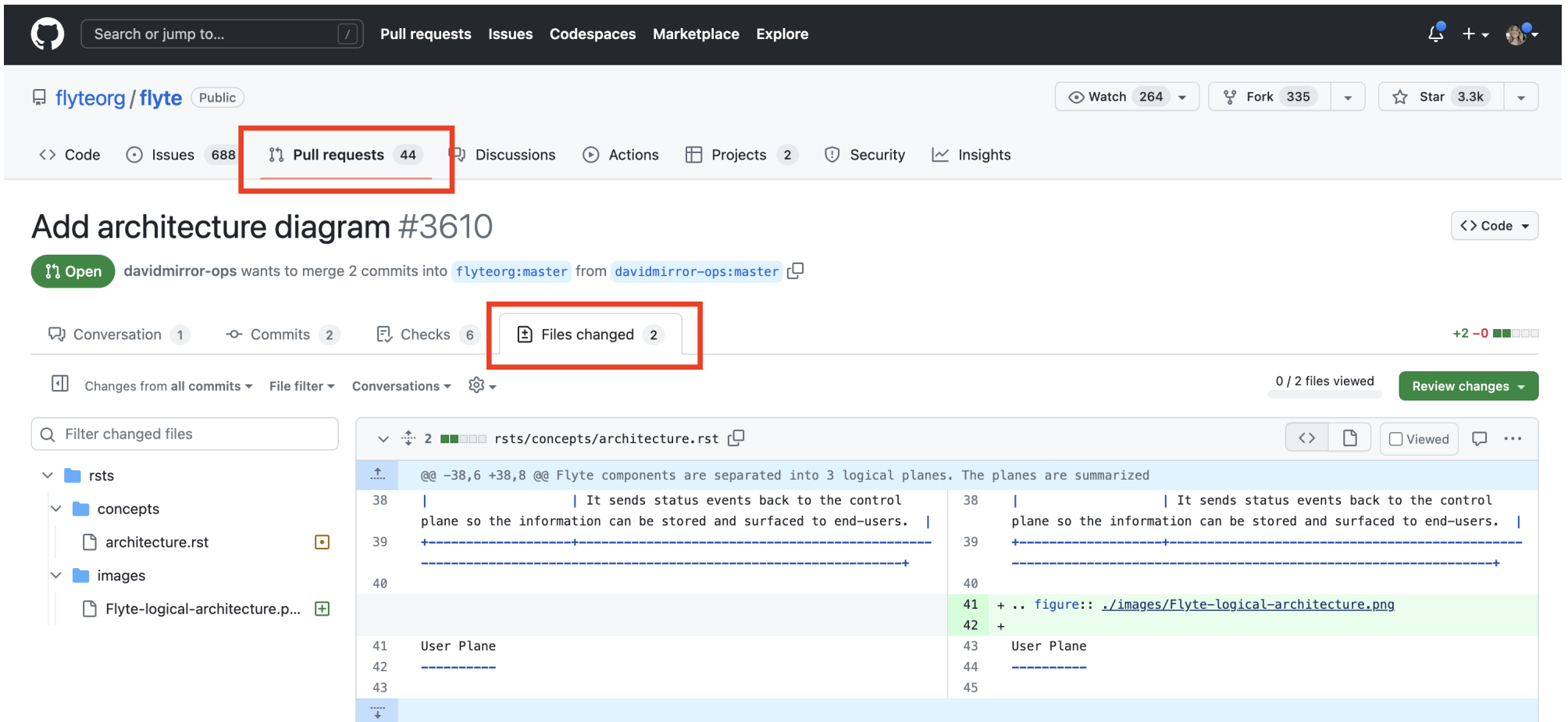

Data gathering
Data versioning
Model monitoring
Orchestration
Data cleaning
Model deployment
Data exploration
Model fitting and evaluation
Model versioning

What I did

analysis of *open source projects* labeled with **MLOps**

Methodology

Github data


Search or jump to...
Pull requests
Issues
Codespaces
Marketplace
Explore
flyteorg / flyte
Public
Watch 264
Fork 335
Star 3.3k
Code
Issues 688
Pull requests 44
Discussions
Actions
Projects 2
Security
Insights
Add architecture diagram #3610
Code
Open
davidmirror-ops wants to merge 2 commits into flyteorg:master from davidmirror-ops:master
Conversation 1
Commits 2
Checks 6
Files changed 2
+2 −0
Changes from all commits
File filter
Conversations
0 / 2 files viewed
Review changes
Filter changed files
rsts
concepts
architecture.rst
images
Flyte-logical-architecture.p...
2 rsts/concepts/architecture.rst
Viewed
@@ -38,6 +38,8 @@ Flyte components are separated into 3 logical planes. The planes are summarized
| | It sends status events back to the control plane so the information can be stored and surfaced to end-users. |
+ .. figure:: ./images/Flyte-logical-architecture.png
+
User Plane
User Plane

Github data

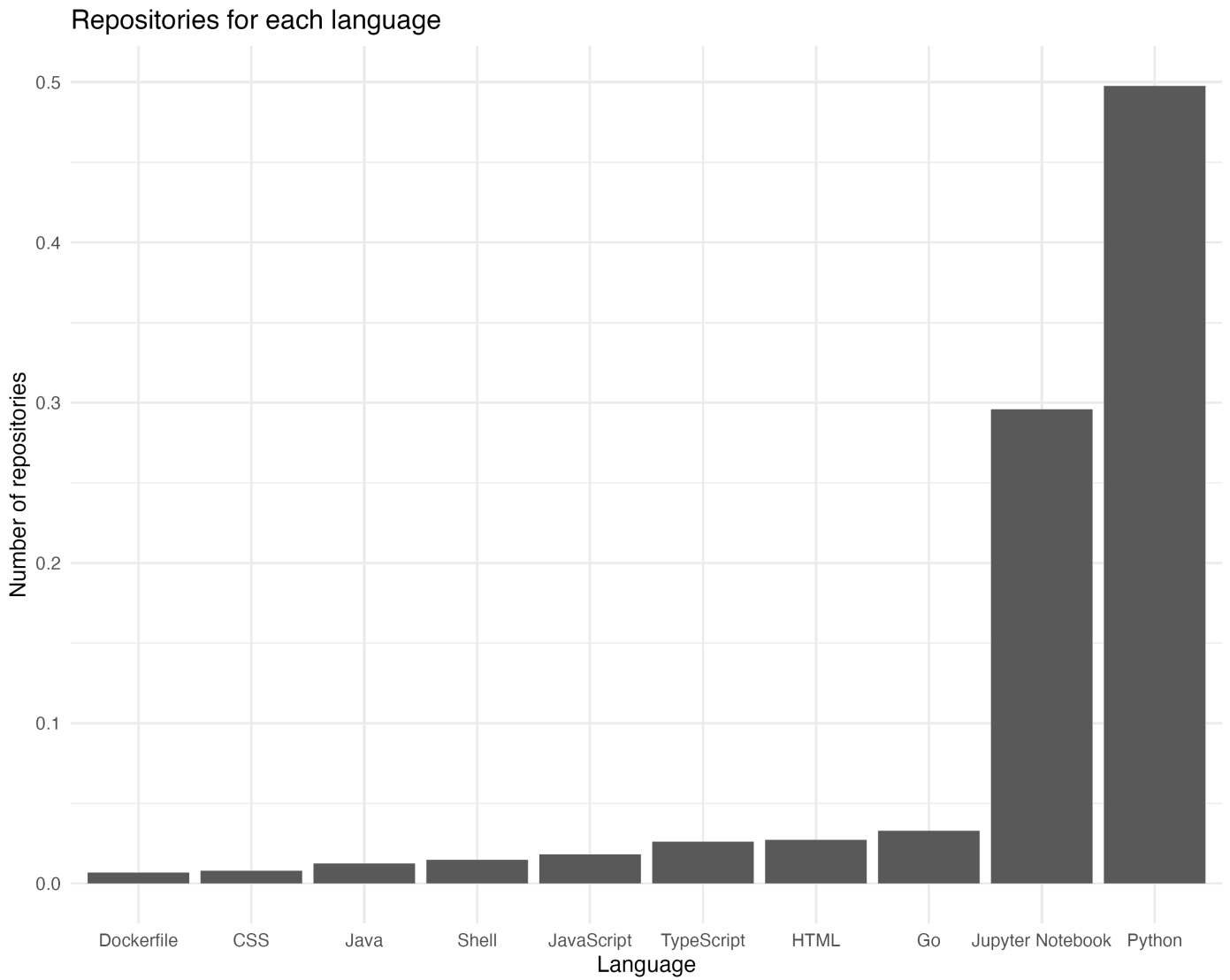
- gathered repositories with `mlops`, `model-management` labels
- programmatic access through REST API to get information

```
gh api \
  -H "Accept: application/vnd.github+json" \
  "/search/repositories?q=topic:mlops&per_page=100&pag
```

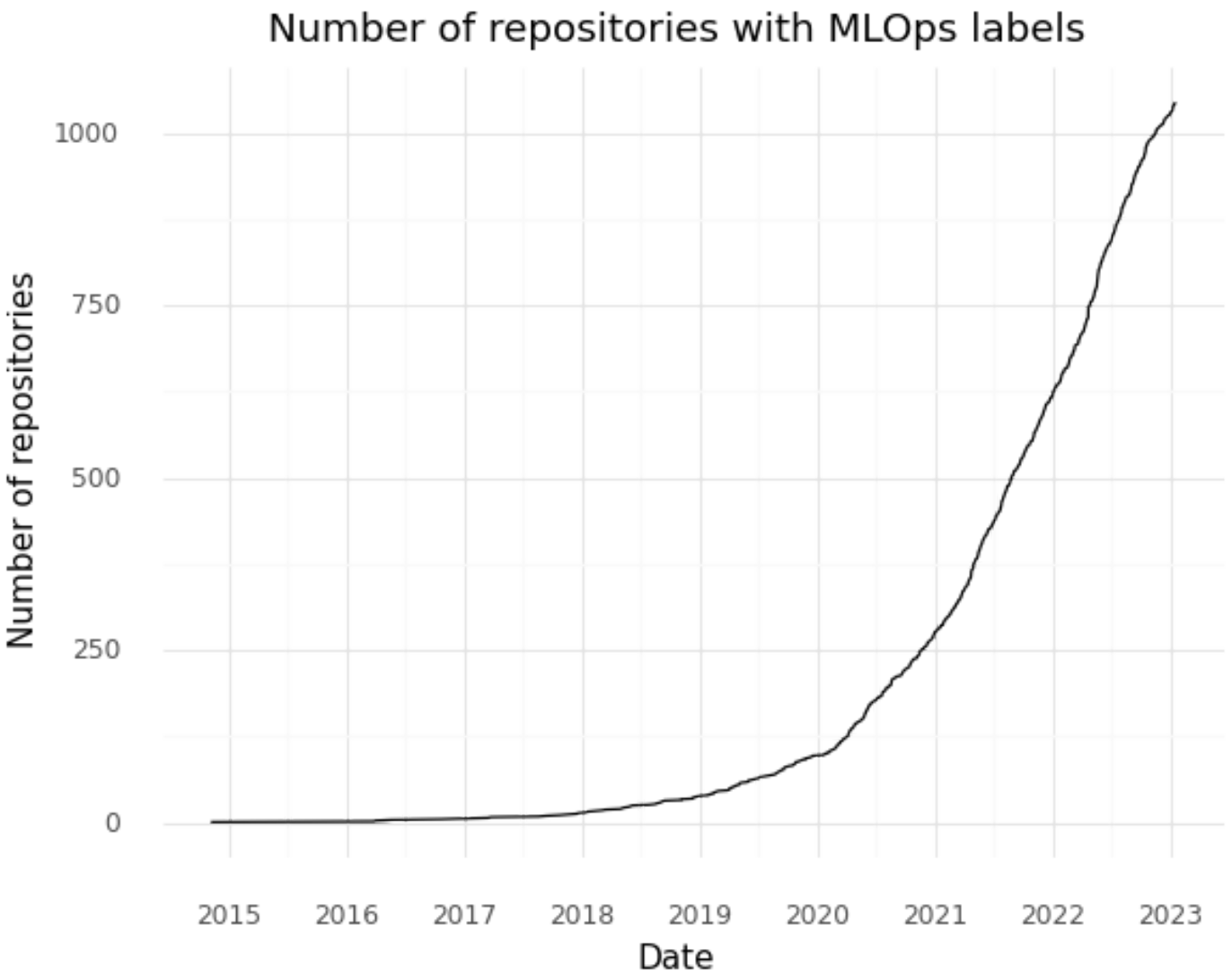
Limitations of data

- all open source users
- based off of self-labeled topics

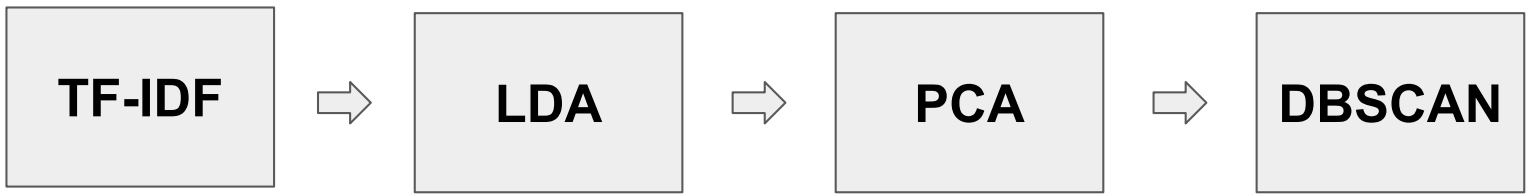
Exploratory data analysis



Exploratory data analysis



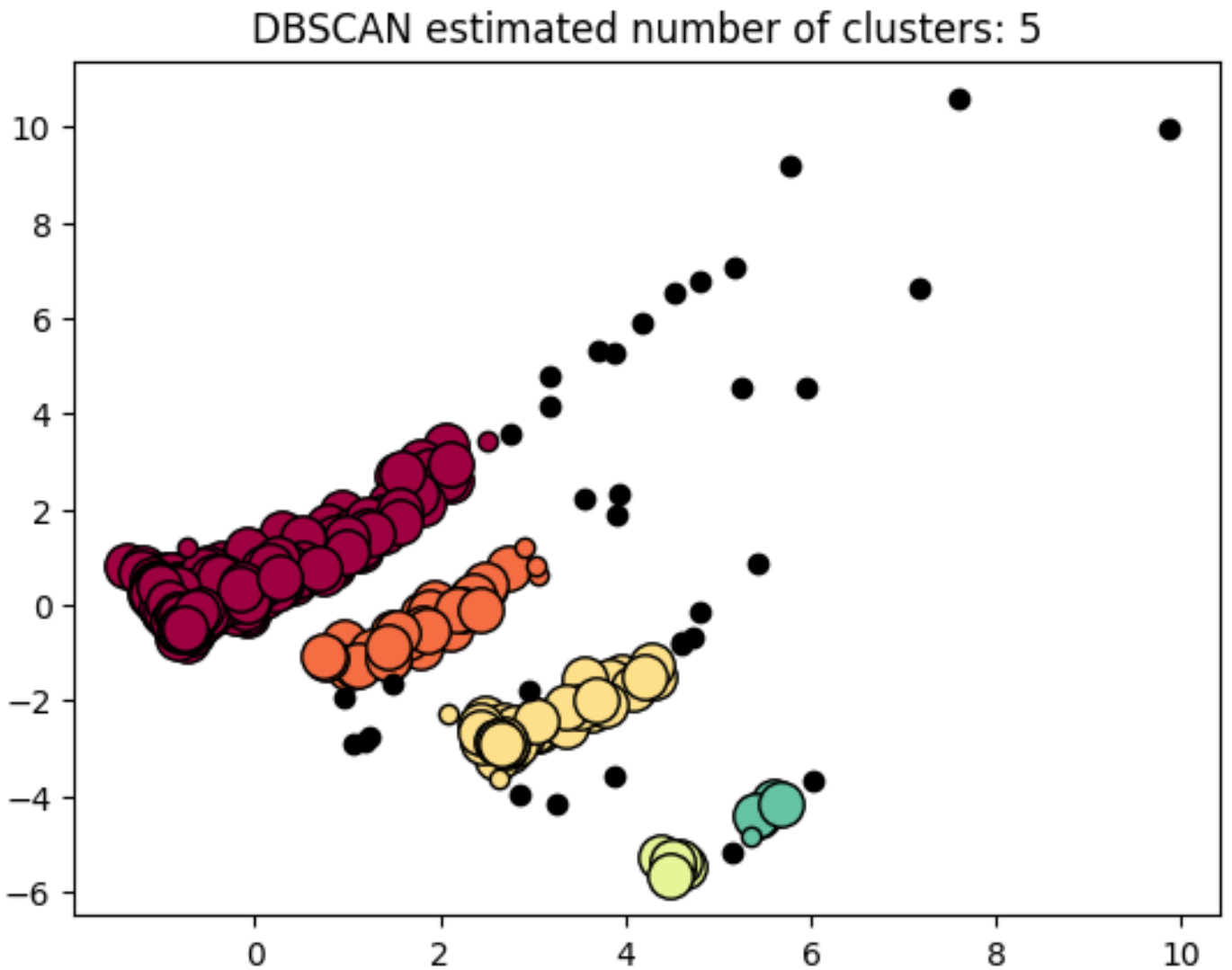
Modeling



Repository
descriptions

stargazers_count
has_issues
has_projects
has_downloads
has_wiki
has_pages
has_discussions
open_issues_count
allow_forking
is_template
age_days
time_since_last_commit
lda_topic

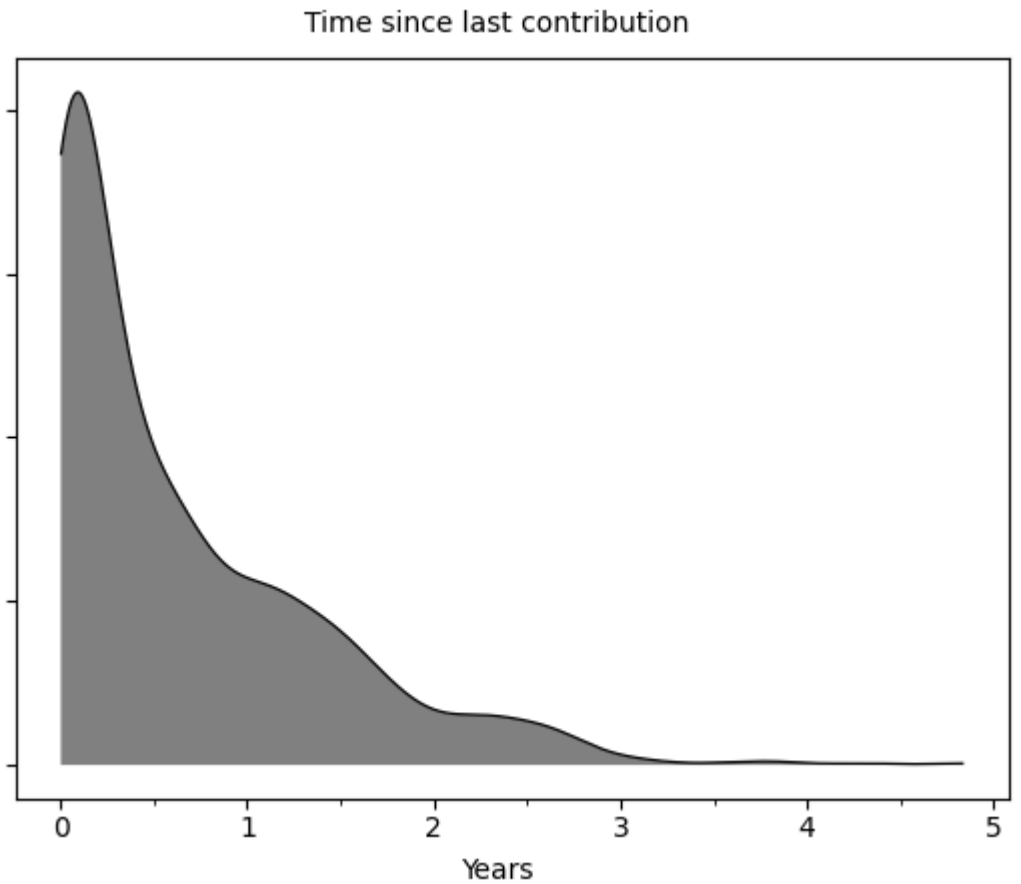
Modeling



Discussion

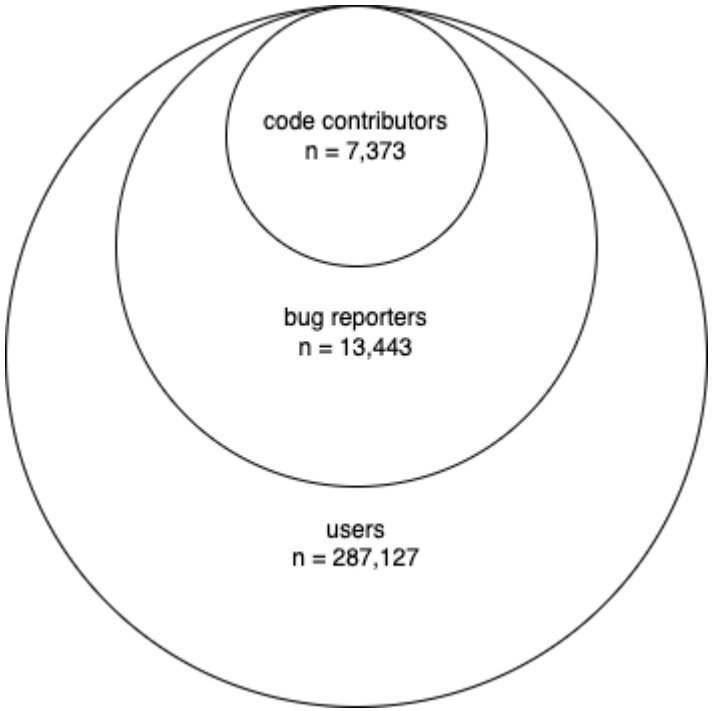
How is this ecosystem growing?

- Recency



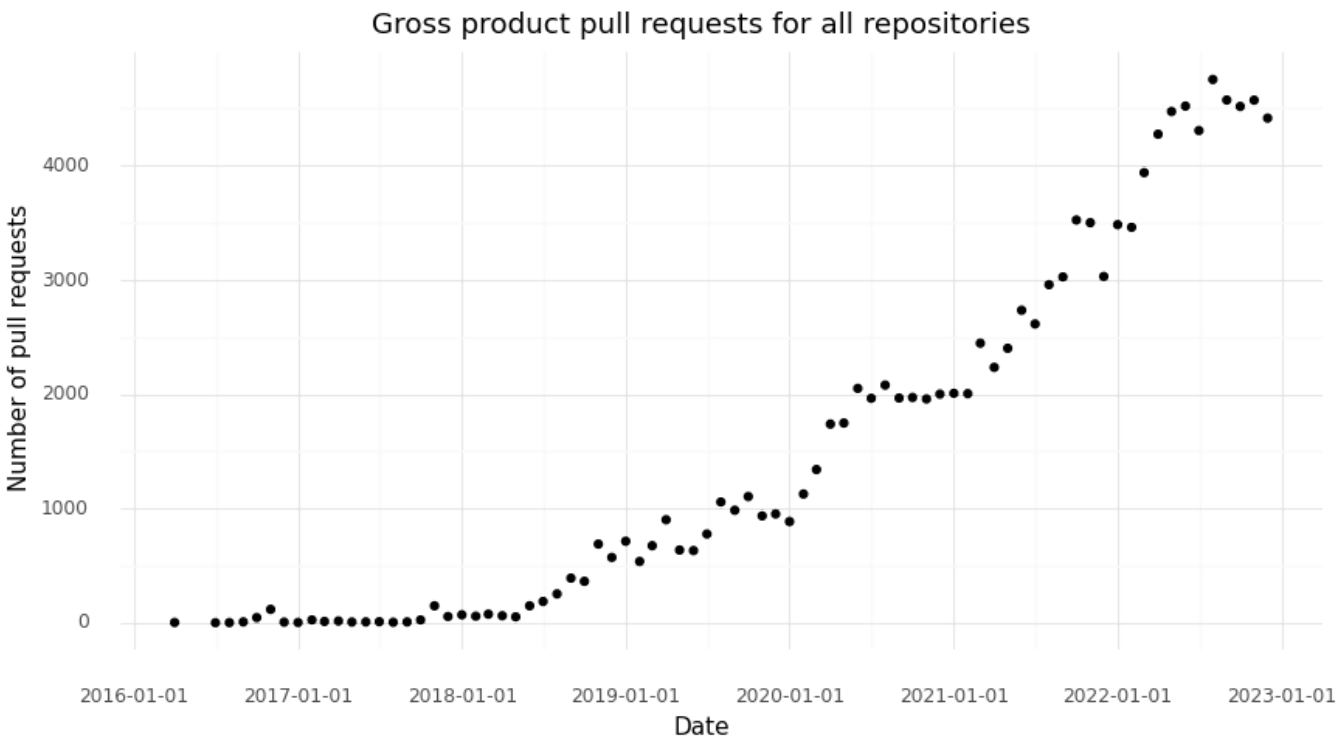
How is this ecosystem growing?

- Recency
- Size

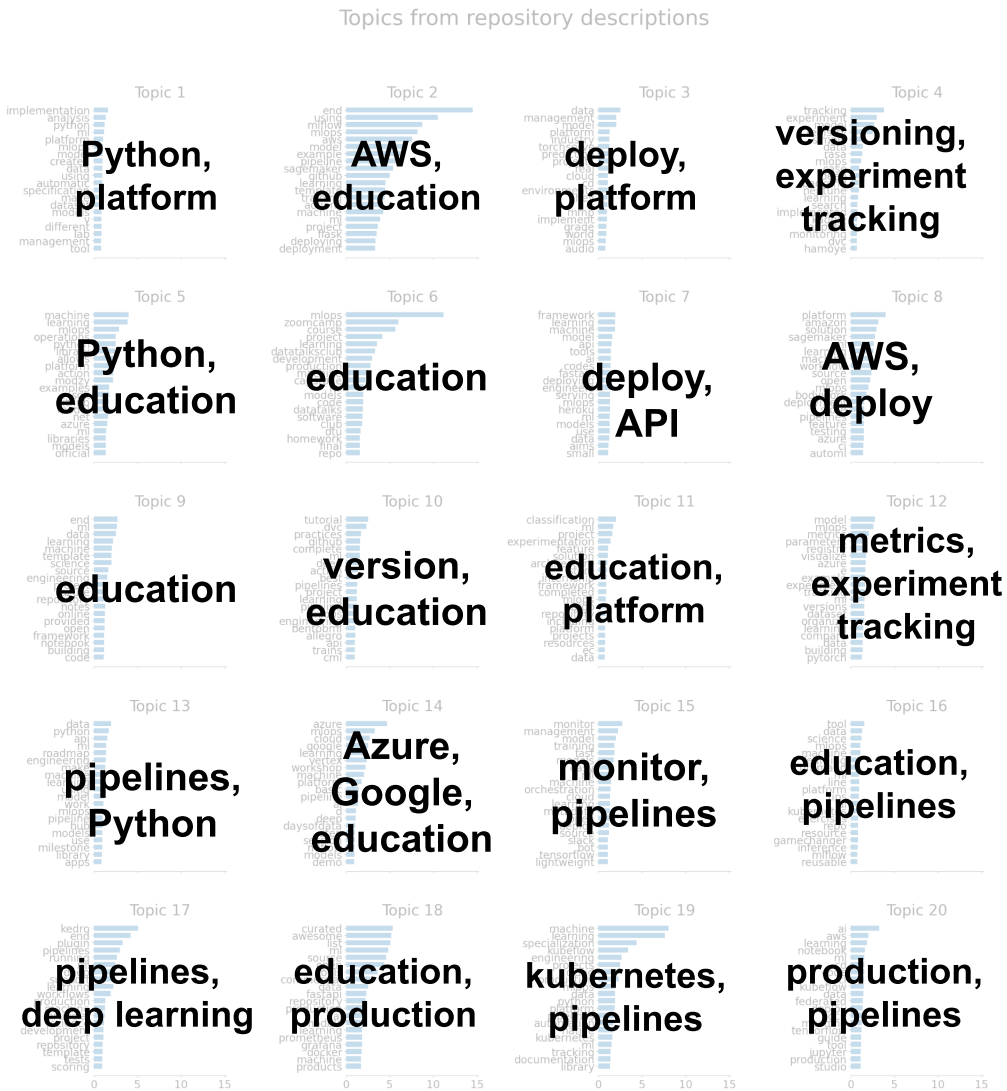


How is this ecosystem growing?

- Recency
- Size
- Contribution rate



What kind of content is being posted?



What does adoption of MLOps look like?

- cloud prevalence
- DataOps vs MLOps

Further work

- repeat this study with all data science and machine learning tags
- opinionated workflow
- survival analysis

Conclusion

- MLOps is a new field built from established practices
- The MLOps open source code base is growing
- Much content is built around learning MLOps
- There may be a need to more strongly segregate DataOps and MLOps
- Cloud infrastructure is highlighted much more often in code than in literature

Thank you all!

And a special thanks to the following:

- Dr. Reinaldo Sanchez-Arias
- Dr. Paniz Abedin
- Dr. Julia Silge
- Dr. Sophie Watson
- Orel Yoshia
- Valentina Colorado
- Noah Zimmerman