Supplementary Materials

Data Flow

1. Take data\tfdd\_data and upload it into bigquery into the schema tfdd and table tfdd. The data is taken from [**On the abandonment and survival of open source projects: An empirical investigation**](https://arxiv.org/pdf/1906.08058.pdf)
2. Take data\ tfdd\_leavers
3. and upload it into bigquery into the schema tfdd and table tfdd\_leavers. The data is taken from [**On the abandonment and survival of open source projects: An empirical investigation**](https://arxiv.org/pdf/1906.08058.pdf)
4. Run code\queries\etl.sql to generate the table valid\_commits
5. Run code\queries\classify\_valid\_commits.sql to generate the table valid\_commits\_classified
6. Run code\queries\describe\_repos.sql to generate the file data\repo\_description.csv
7. Run code\queries\tfdd\_statistics to create table tfdd\_raw\_statistics
8. code\queries\tfdd\_diff\_statistics to create table tfdd\_diff\_statistics
9. Run code\queries\tfdd\_leavers\_statistics to create table tfdd\_ leavers\_raw\_statistics
10. code\queries\tfdd\_ leavers\_diff\_statistics to create table tfdd\_ leavers\_diff\_statistics and generate tfdd\_leavers\_diff\_stats\_at\_least\_9

# Code

## Queries

The queries are run on the [GitHub database available in Google's BigQuery](https://bigquery.cloud.google.com/dataset/bigquery-public-data:github_repos) infrastructure.

Most queries were run directly on the GitHub tables.

Some flows included intermediate tables. These table are stored to a schema that is unique per user. When running these queries, the schema name should be named to the schema name given to your user.

* ETL - Extract, Transfer and Load: Moving the data to a convenient structure
* classify\_valid\_commits - label all valid commits with linguistic models

# Data

* tfdd\_data – a list of truck factor developer disattachment. Data is based on the paper <FILL IN>. Contains evented that all the truck factor developer left a project (the last commit of the last developer).
* tfdd\_leavers – A close in spirit data set to tfdd\_data. Mark the event of any truck factor developer leaving the project (the last commit of a truck factor developer). . Data is based on the paper <FILL IN>.