Preliminary analysis of comp-users-forum

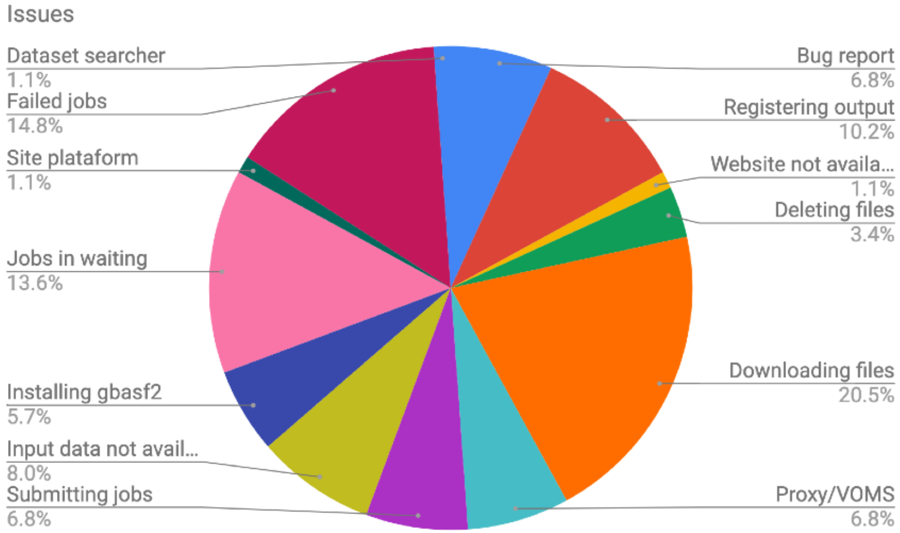
D.Jaffe, 10 Nov 2021

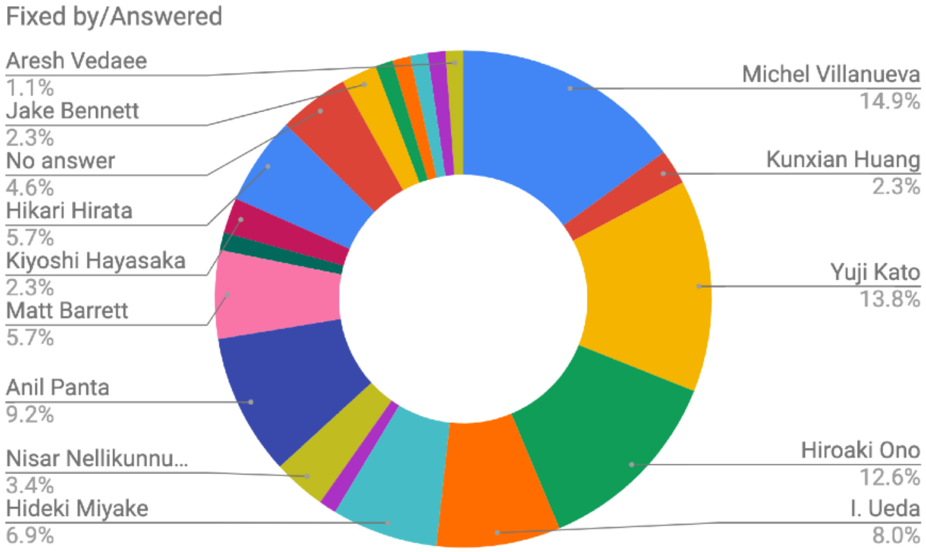
U.Miss/BNL meeting

Goals

- Identify and quantify user issues.
- Identify mitigation possibilities with resources, effort, methodology and/or workflow
- Try to automate Michel's by-hand accounting work (example from June 2021 B2GM)

  - Most common issues:
    - Accessing files in the SEs (downloading fail, input data is not available).
    - Jobs in waiting status too long.



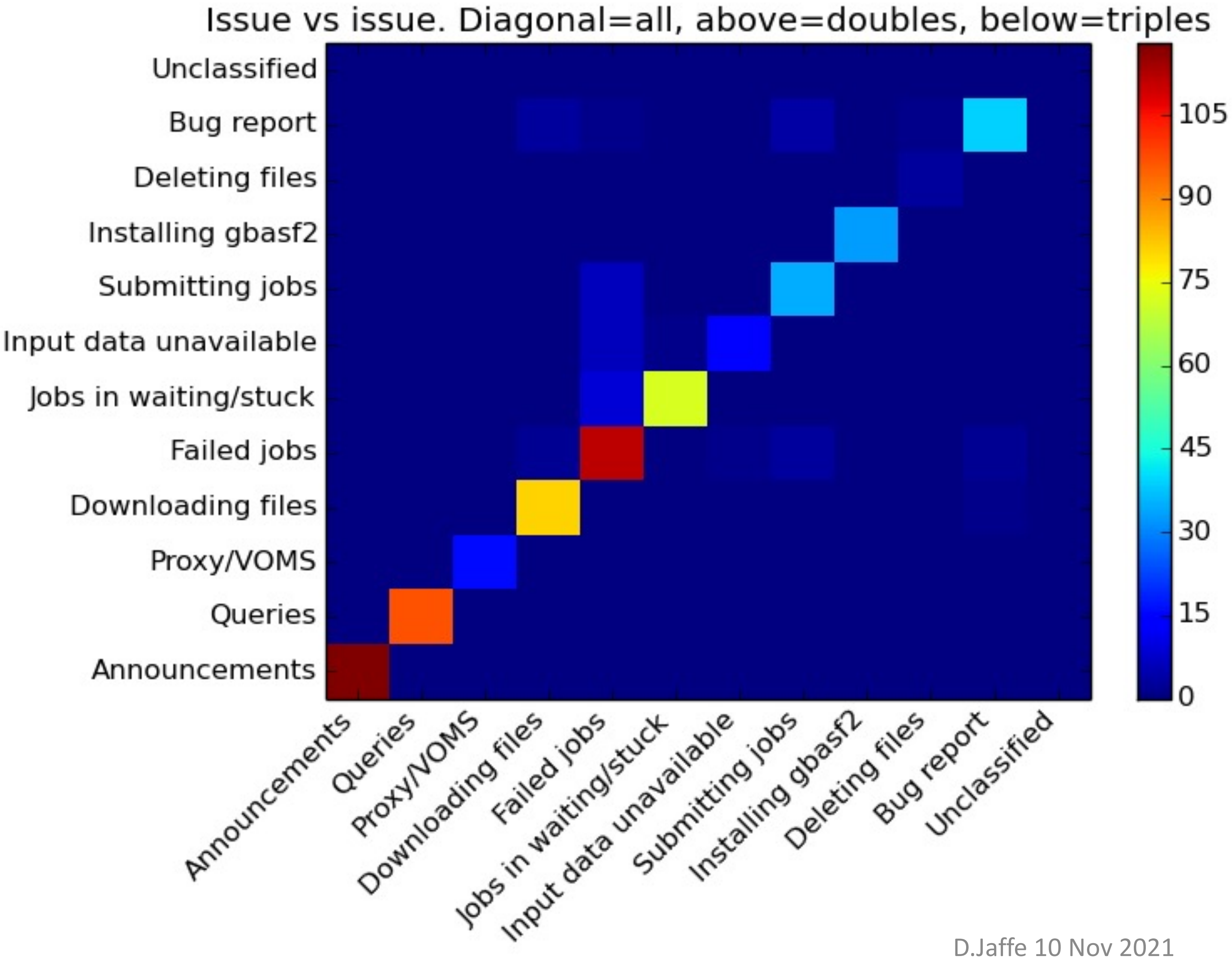


- Many people contributed to solve issues and questions. Thanks to all!
  - Expert shifters reply timely to users.
- We have **4 messages without answer**.

Analysis technique

- Take guidance from Michel classification of issues
- Start from zip file of comp-users-forum archive from Feb 2017 – Oct 2021 provided by Hara-san.
- Reconstruct threads using In-Reply-To, References and Subject fields. Less than 10% of threads improperly reconstructed. (There is no thread information provided with archive zip.)
- Classify threads into issues using Subject first, then message content.
- Track reporters (initial submitter) and responders (first response in thread).
- Determine thread resolution time (=t(last msg in thread) – t(start of thread)

Some results (5 Nov 2021 analysis, 20211105T135449)



Issue classification

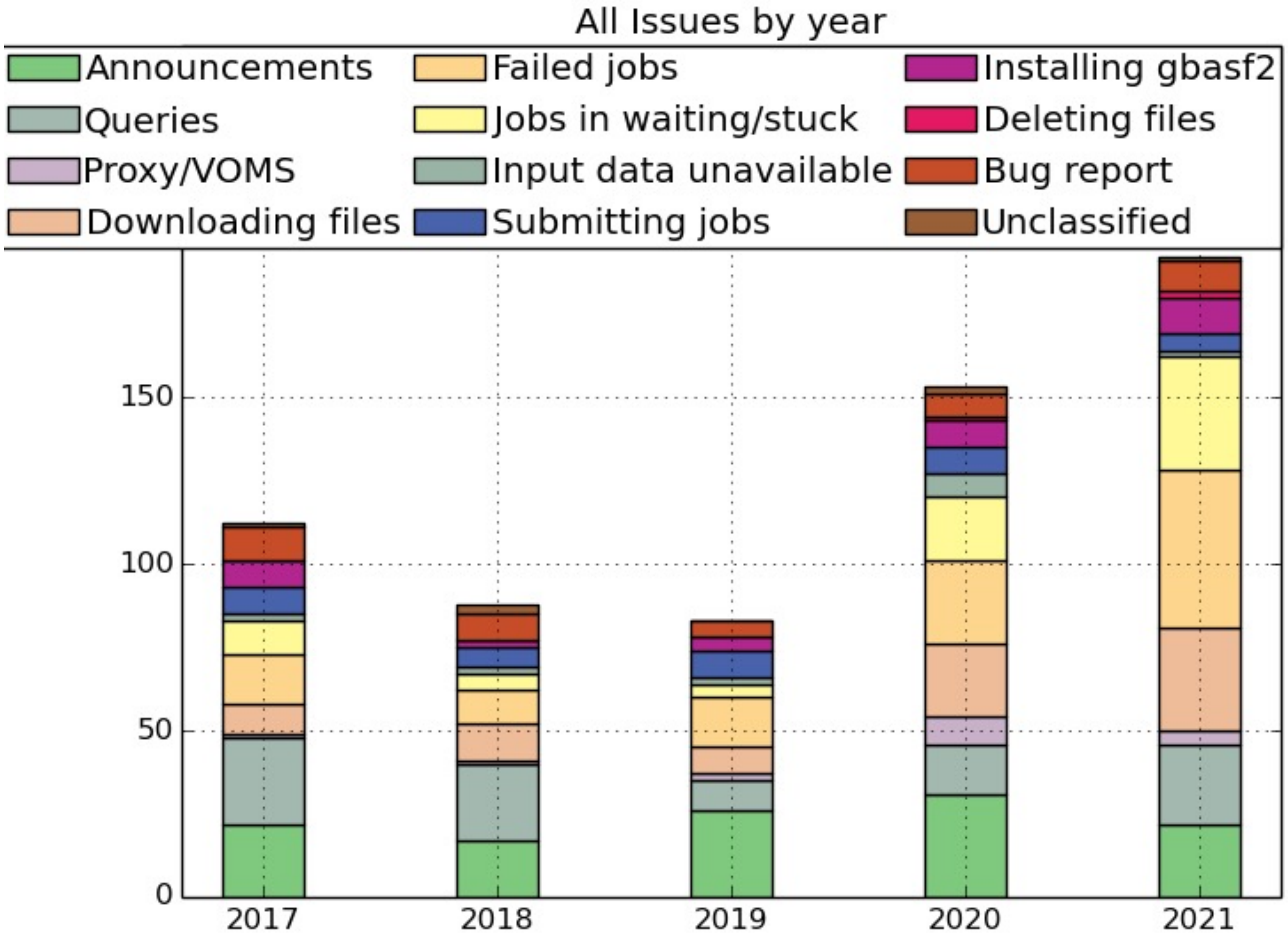
- 2165 messages classified into 588 threads
- 26 threads classified under 2 issues
- 7 threads classified under 3 issues
- 7 threads 'Unclassified'

Single issue classification required for

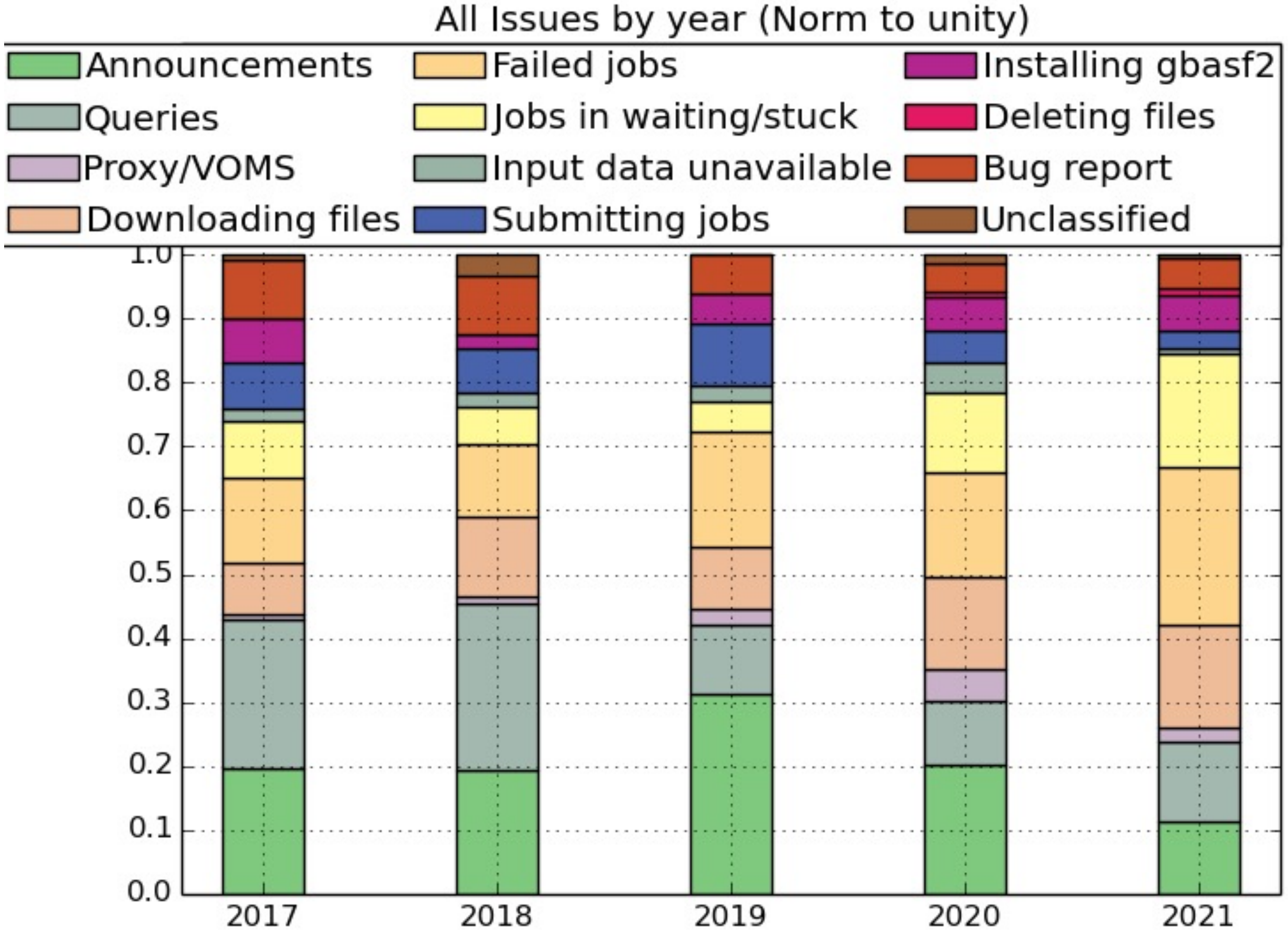
- Announcements
- Queries
- Proxy/VOMS

Other Issues

- Downloading files
- Failed jobs
- Jobs in waiting/stuck
- Input data unvailable
- Submitting jobs
- Installing gbasf2
- Deleting files
- Bug report

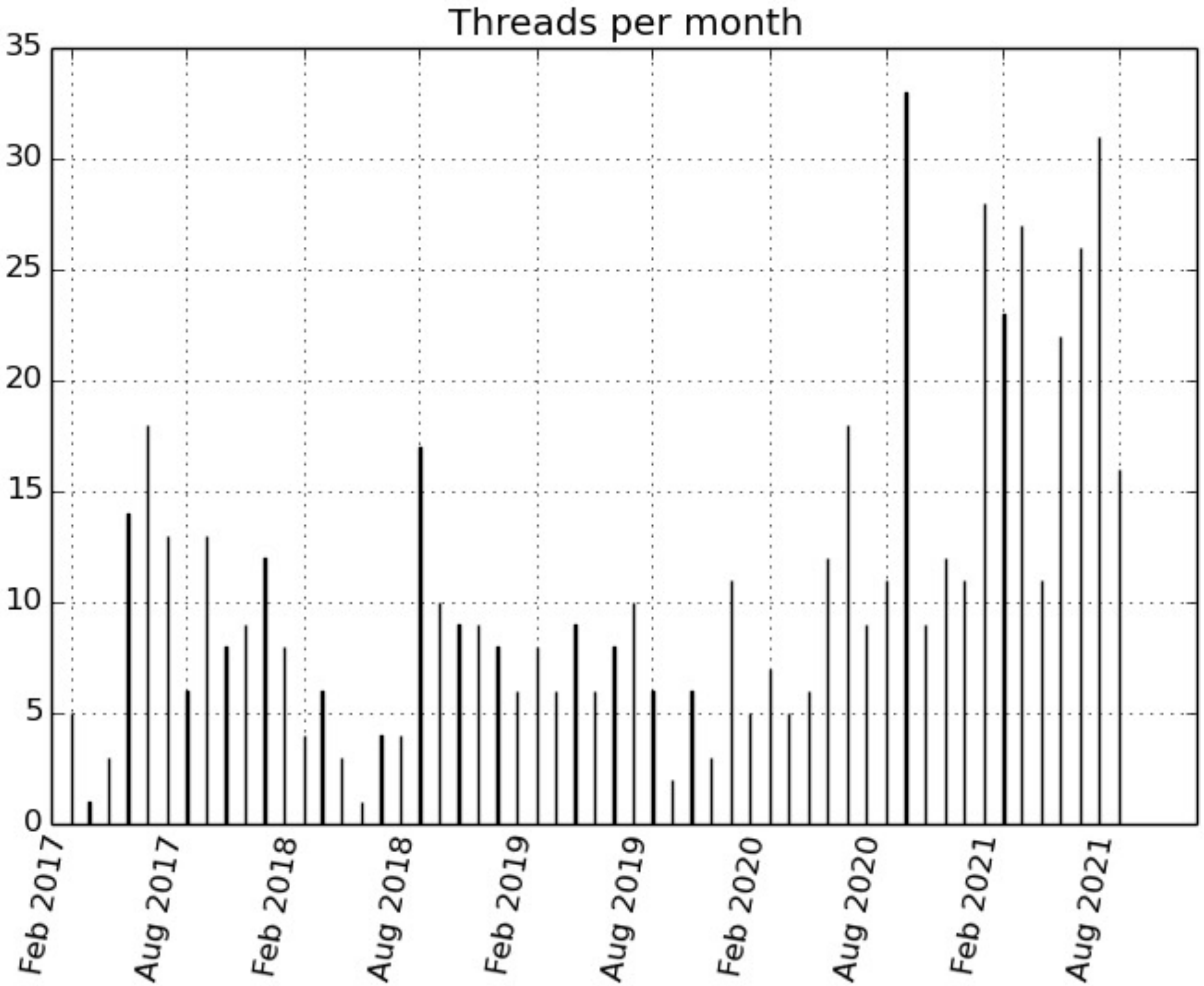


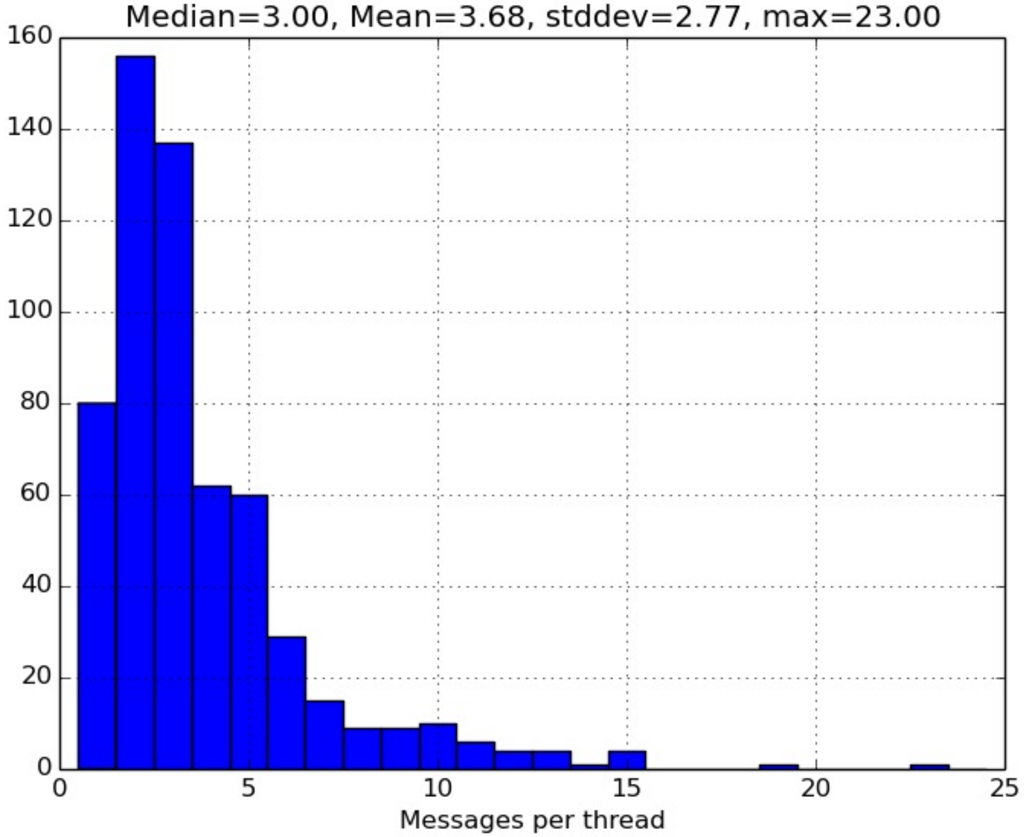
- Roughly constant number of announcements and queries.
- Job failures and file download issues increasing with time (probably due to available data)



Same as previous, but total issues by year are normalized to unity to judge relative frequency.

Threads/month and messages/thread

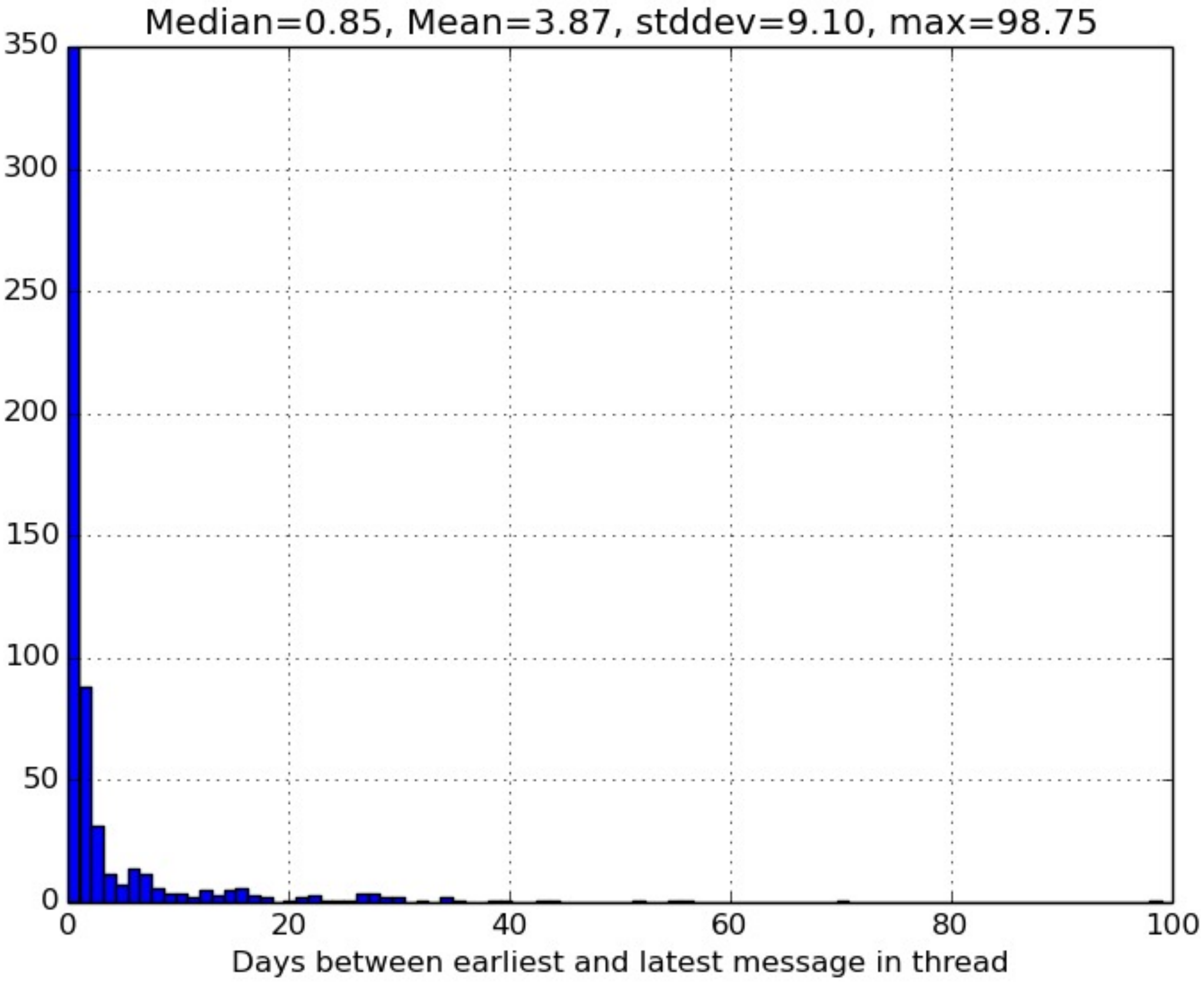




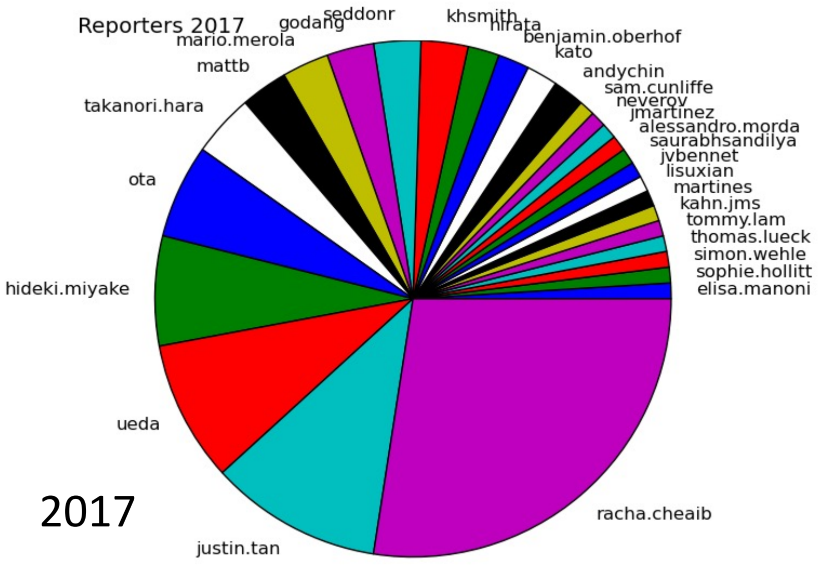
Have not checked to see if threads/month or messages/thread correlated with events (e.g. Rucio migration, gbasf2 release)

Issue resolution time

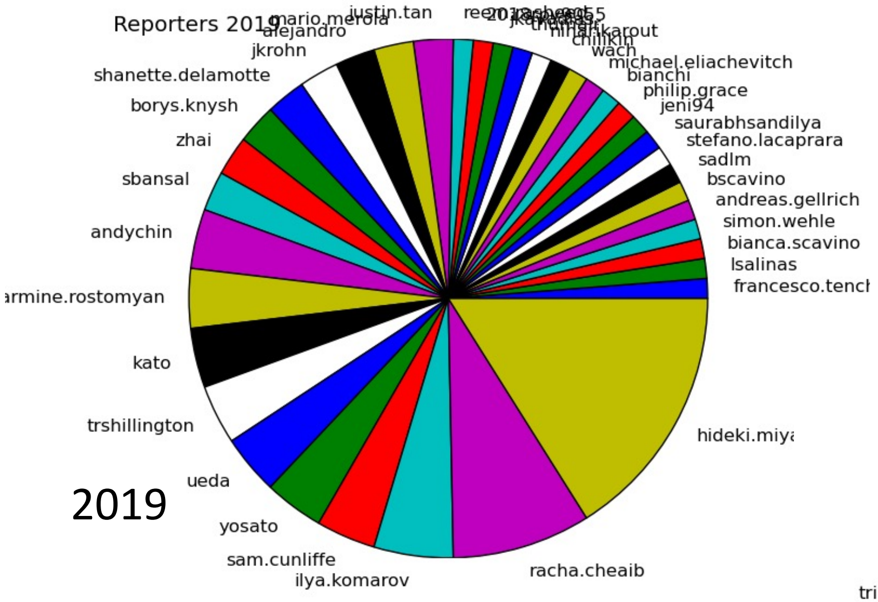
- Thread misidentification is responsible for 99 day ‘resolution time’. This highlights the need to improve thread identification.
- Message time not corrected for time zone
- More than half of all issues resolved in <1 day
- Very few issues unresolved for more than ~10 days.



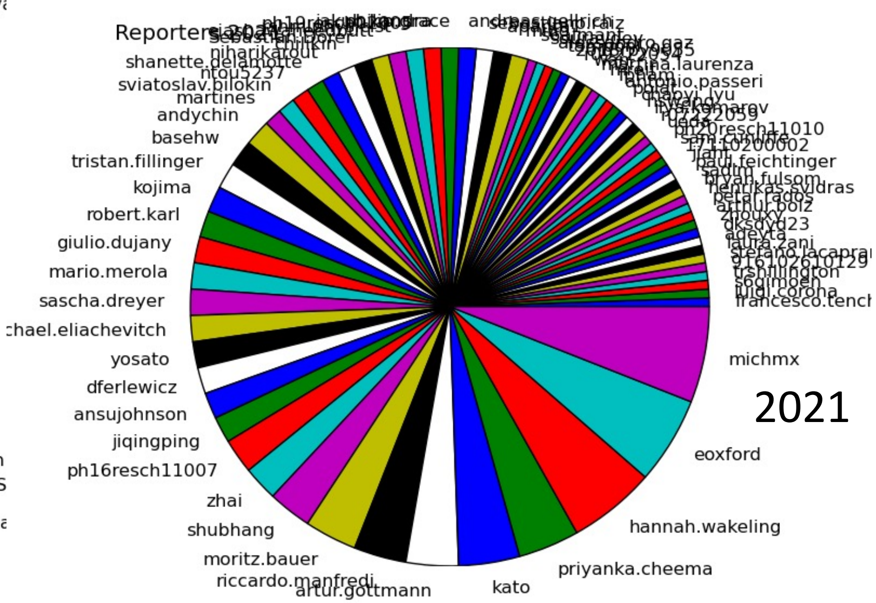
Reporters by year



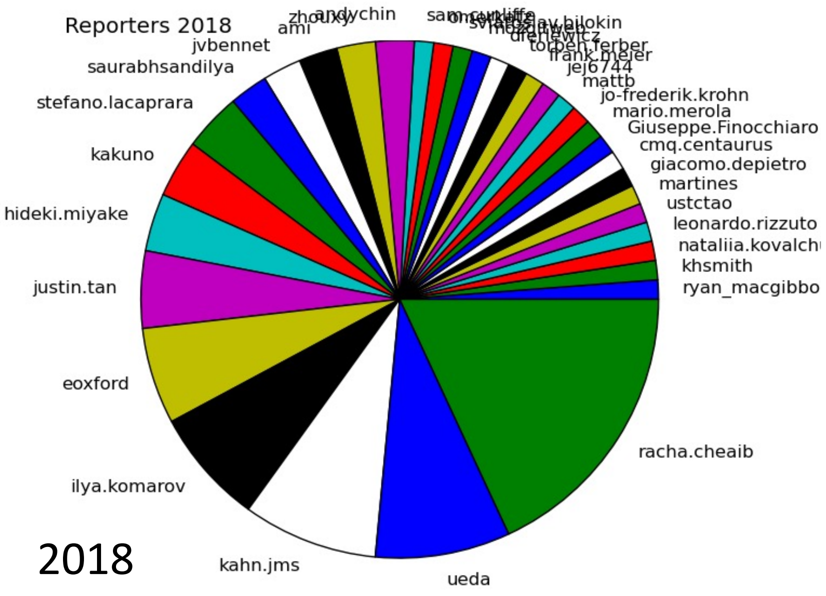
2017



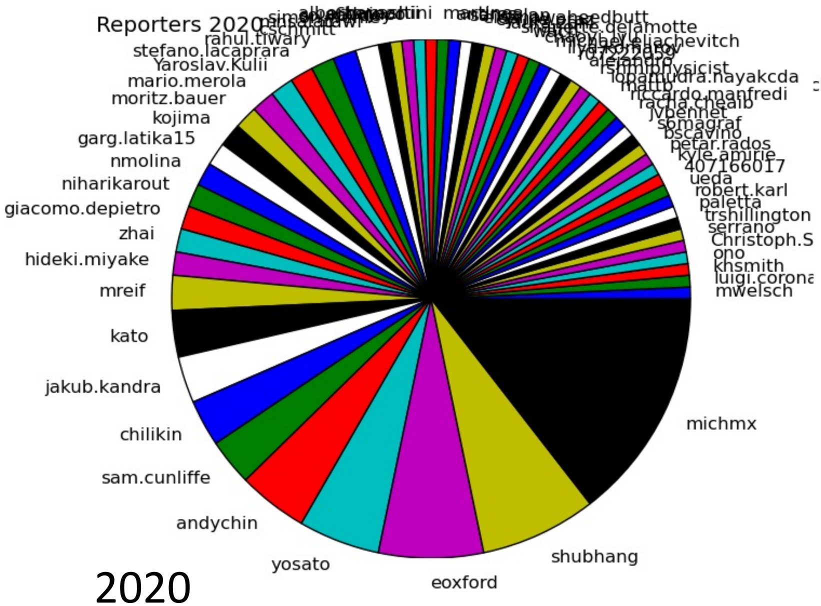
2019



2021



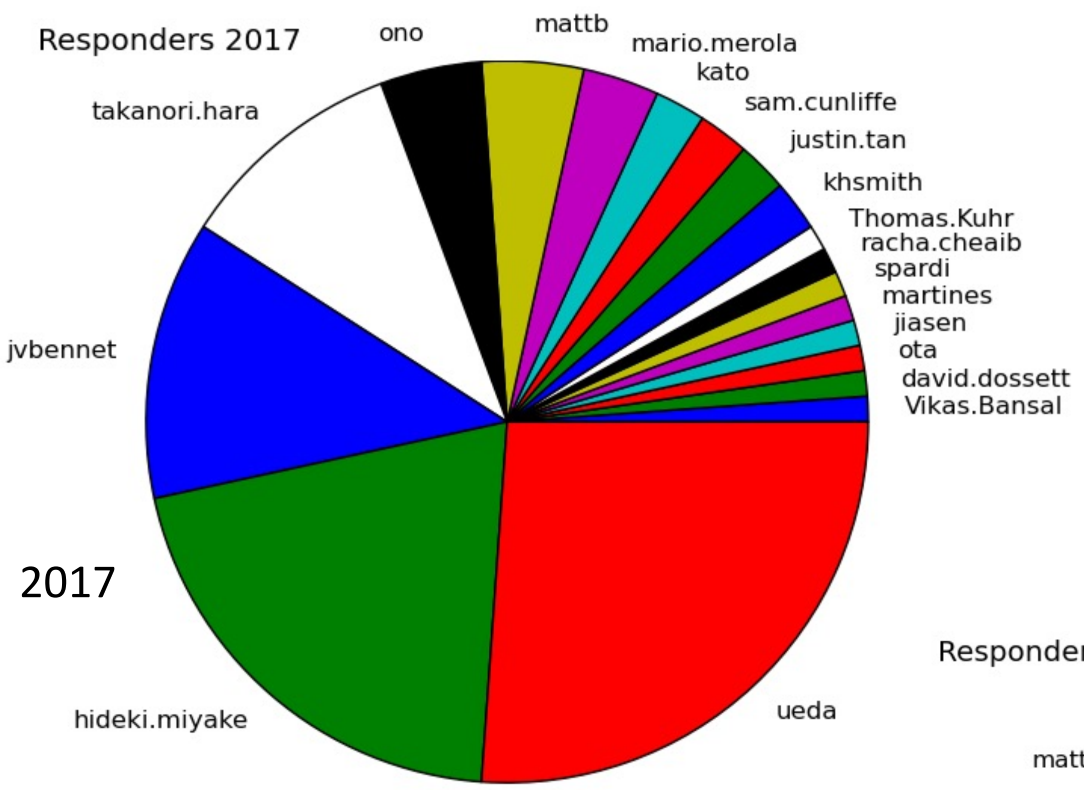
2018



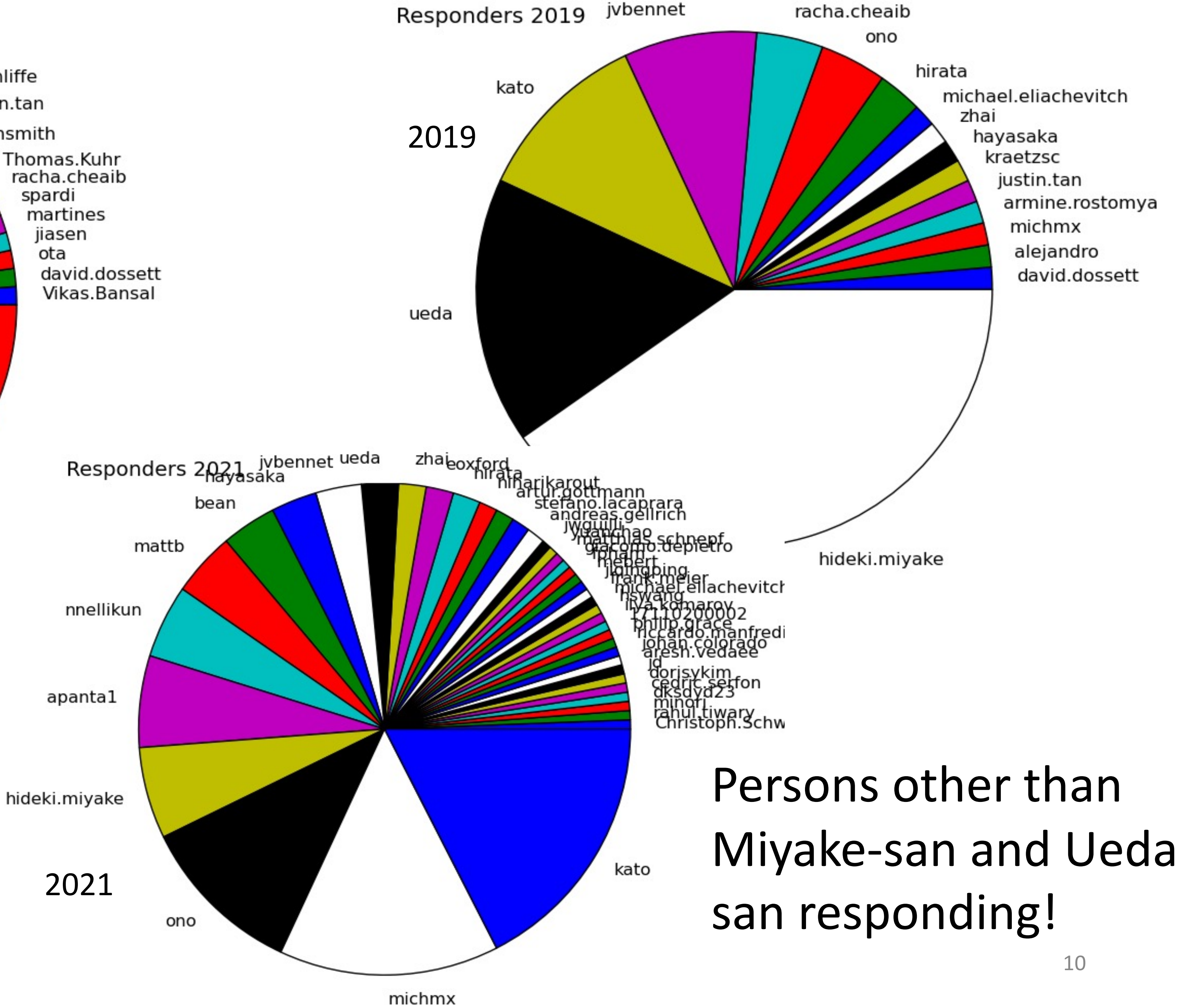
2020

Less dominance by single reporter vs time

Responders by year (2017,19,21)



2017

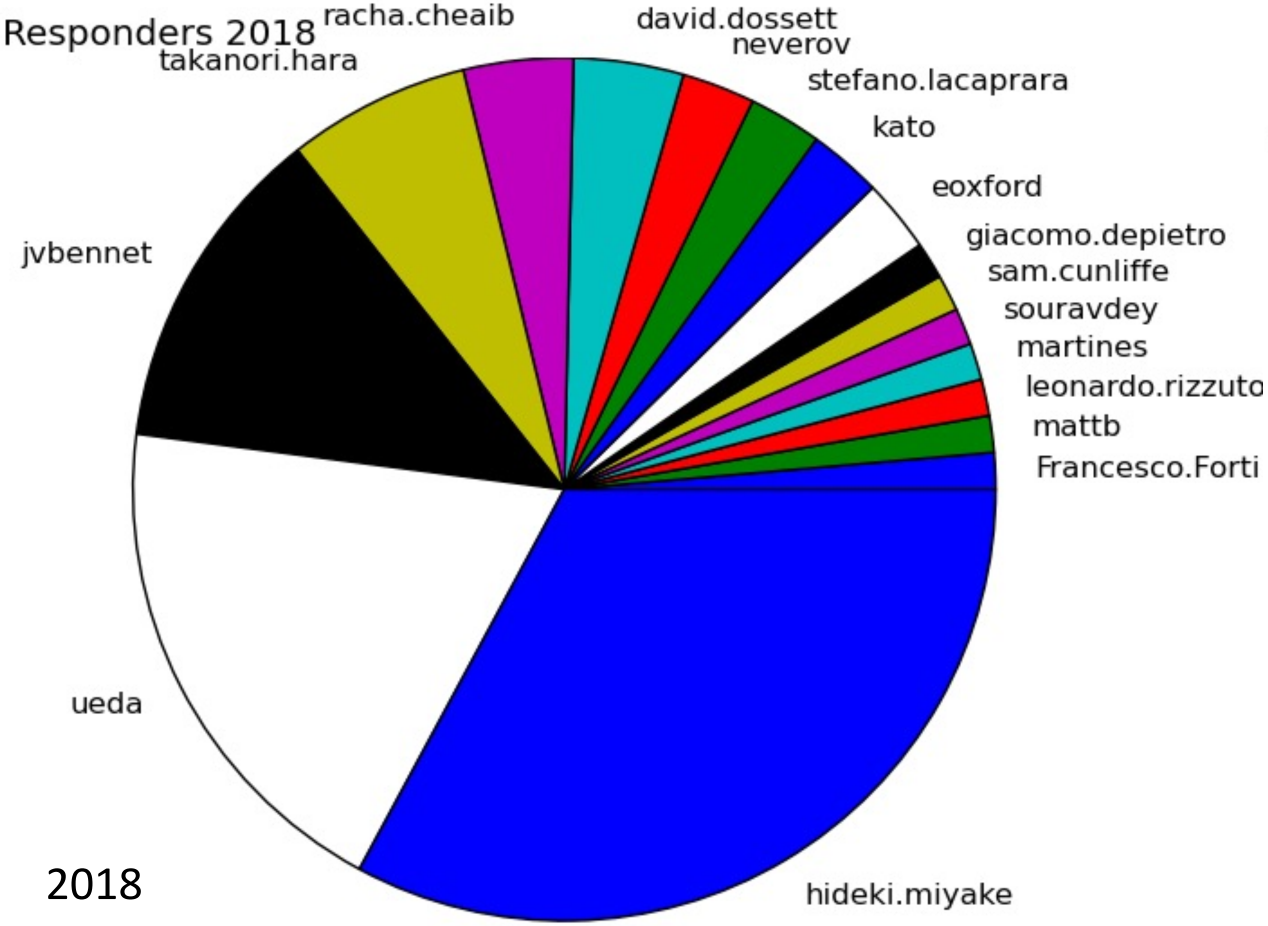


2019

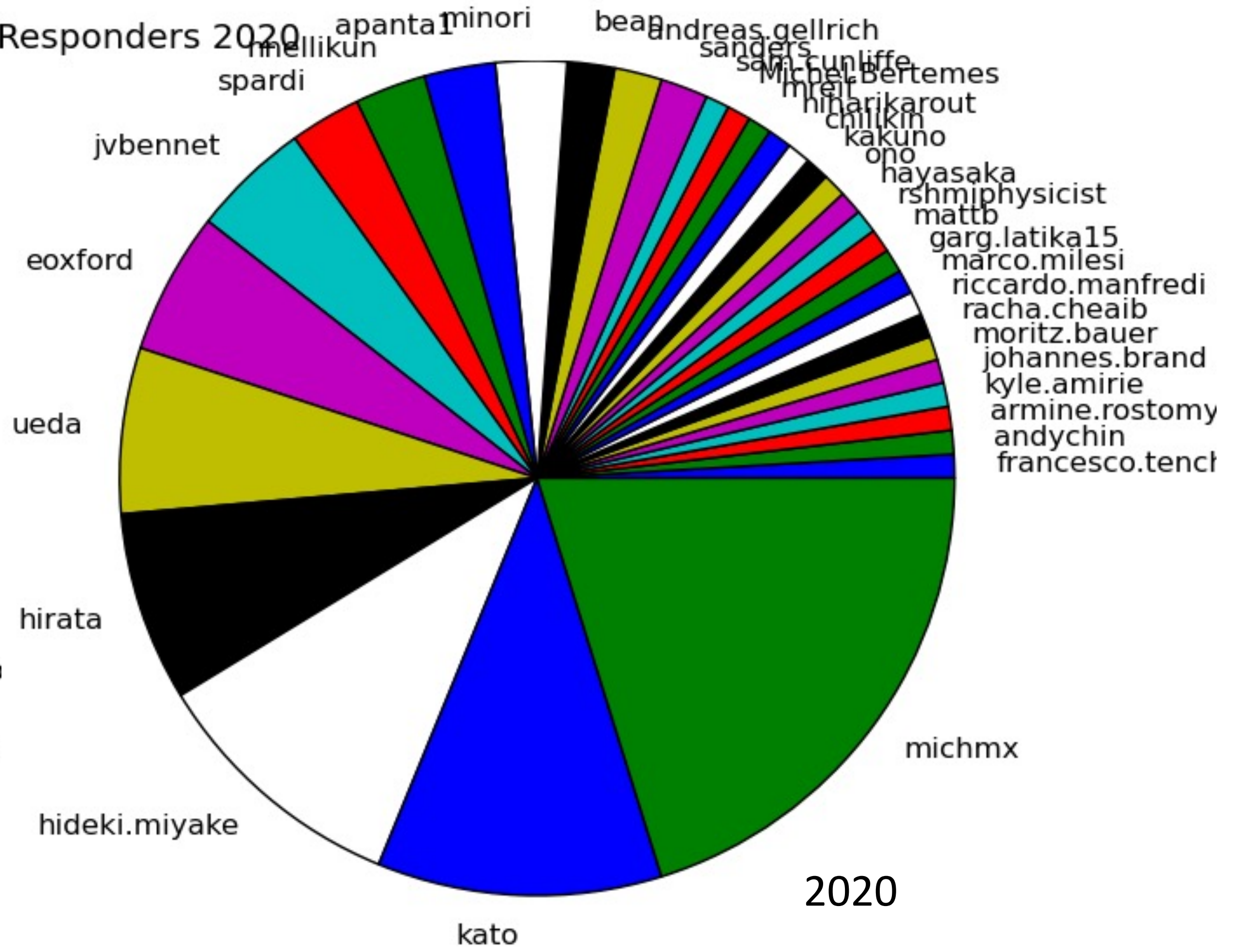
2021

Persons other than Miyake-san and Ueda-san responding!

Responder/yr 2018, 2020



2018



2020

Summary

- Trends identified by Michel are also found by automated methods
- Thread identification still needs improvement. It would be good if thread identification by mailing list could be imported. I have not thoroughly investigated this possibility.
- Would like feedback and suggestions on statistics to report and how to classify and correlate issues
- Work in progress:
  - Correlation of issues and individual grid sites
  - Rucio-related issues

python2.7 code in github