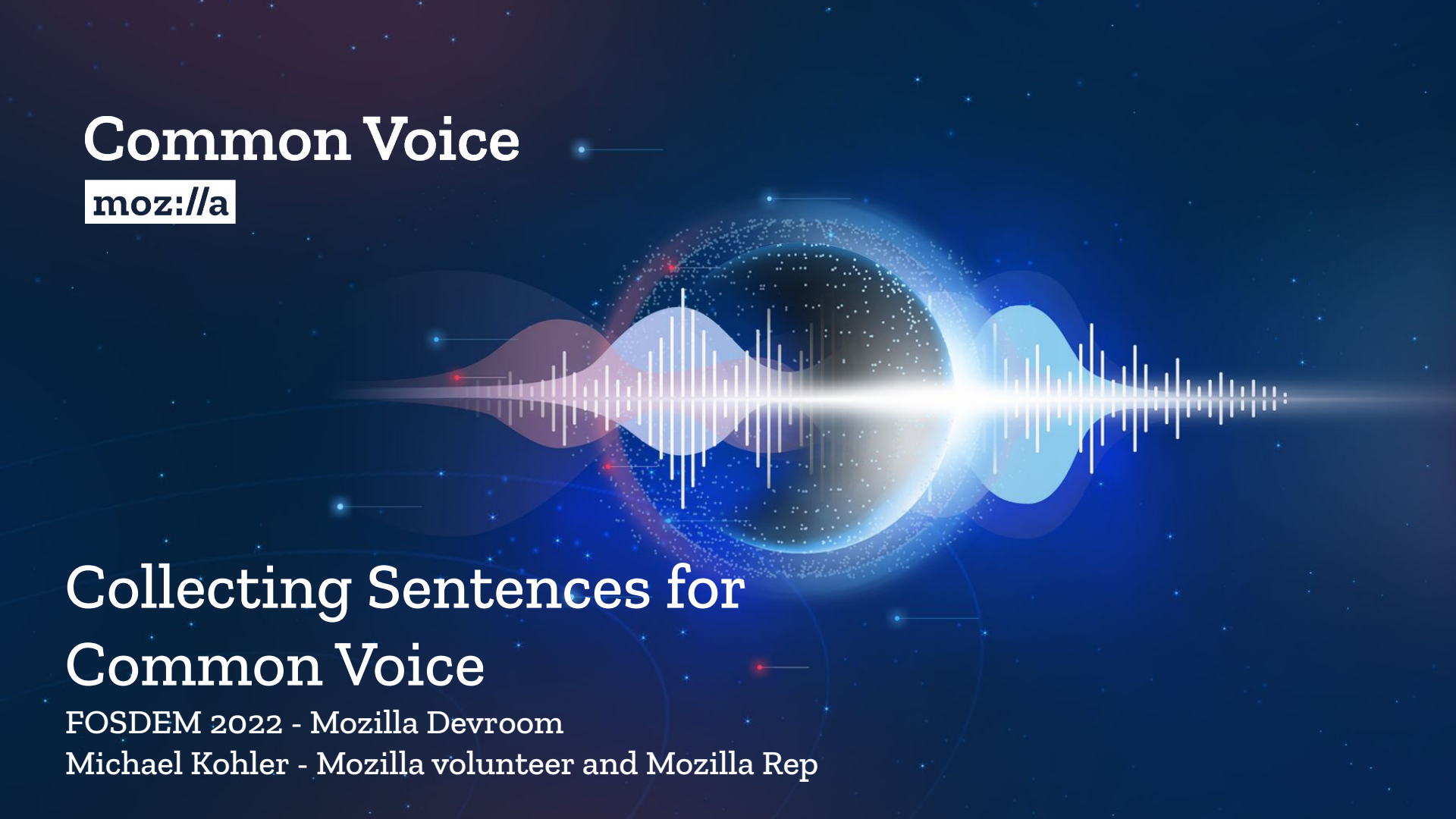
Common Voice

moz://a



Collecting Sentences for Common Voice

FOSDEM 2022 - Mozilla Devroom

Michael Kohler - Mozilla volunteer and Mozilla Rep

What is Common Voice?

- https://commonvoice.mozilla.org
- Project led by Mozilla Foundation
- Project to help make voice recognition open and accessible
- Dataset of recordings and sentences published under CC0 license

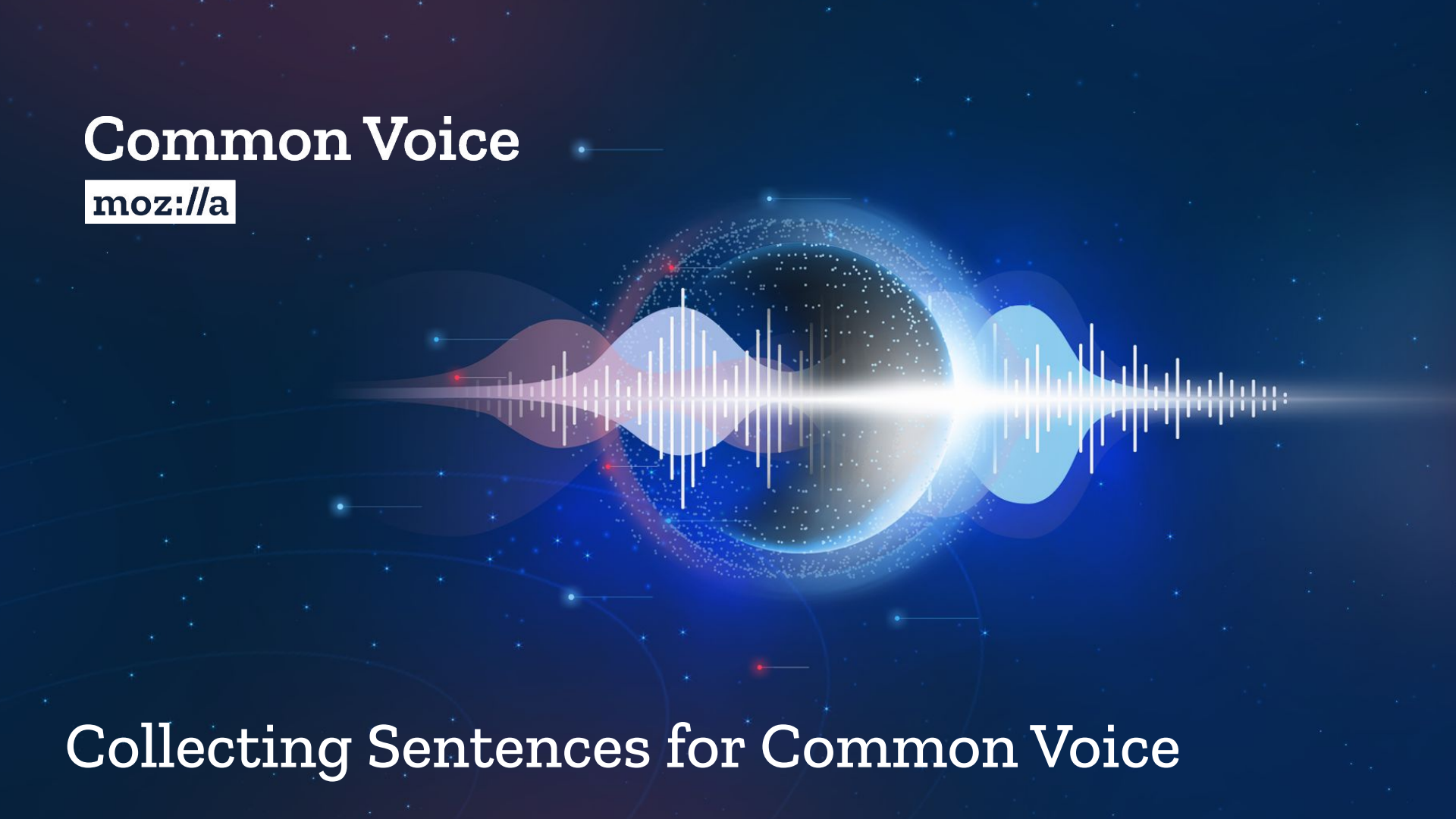
- 159 languages
- 88 languages open for recordings
- Not only the “big” languages!

Dataset

- Released every 6 months
- See https://commonvoice.mozilla.org/datasets for statistics for each language and download

Common Voice

moz://a



Collecting Sentences for Common Voice

Sentence Collector

- Adding sentences for others to record
- Must be free of copyright / CC0 - regularly exported to Common Voice

In this section:

- Demo
- Architecture
- Export

Sentence Collector: https://commonvoice.mozilla.org/sentence-collector/
GitHub: https://github.com/common-voice/sentence-collector/

Sentence Extractor

- Extracting max. 3 sentences per article from Wikipedia
- Other sources available such as WikiSource
- Based on Rule definitions
- Based on GitHub Actions

In this section:

- Rule file structure
- Architecture
- Export

Github: https://github.com/Common-Voice/cv-sentence-extractor

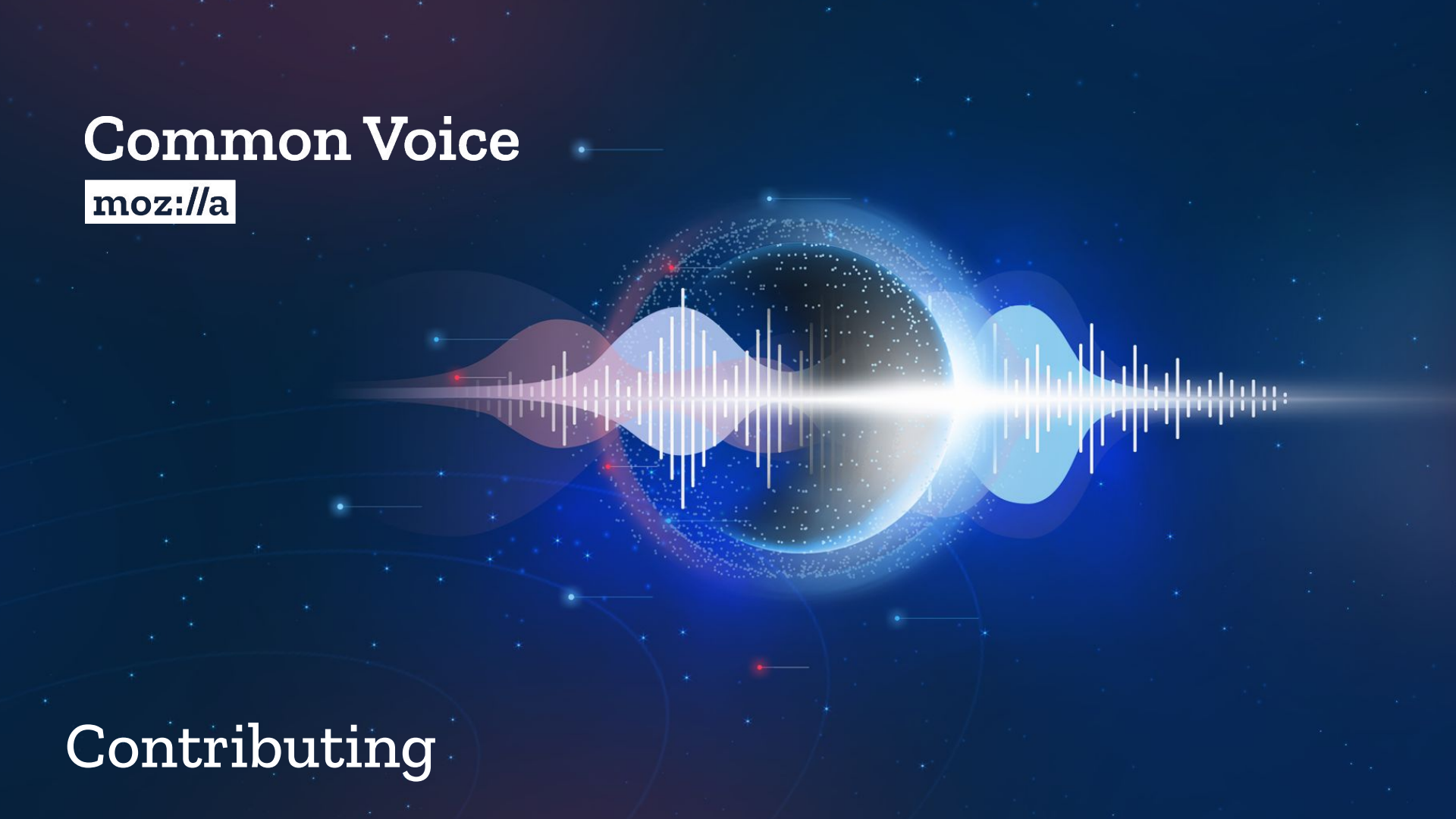
Bulk Uploads

- If upload through Sentence Collector not suitable (larger sets)
- Direct PR to the Common Voice repo
- Review of a sample by multiple reviewers

More info in the playbook:
https://common-voice.github.io/community-playbook/sub_pages/text.html

Common Voice

moz://a



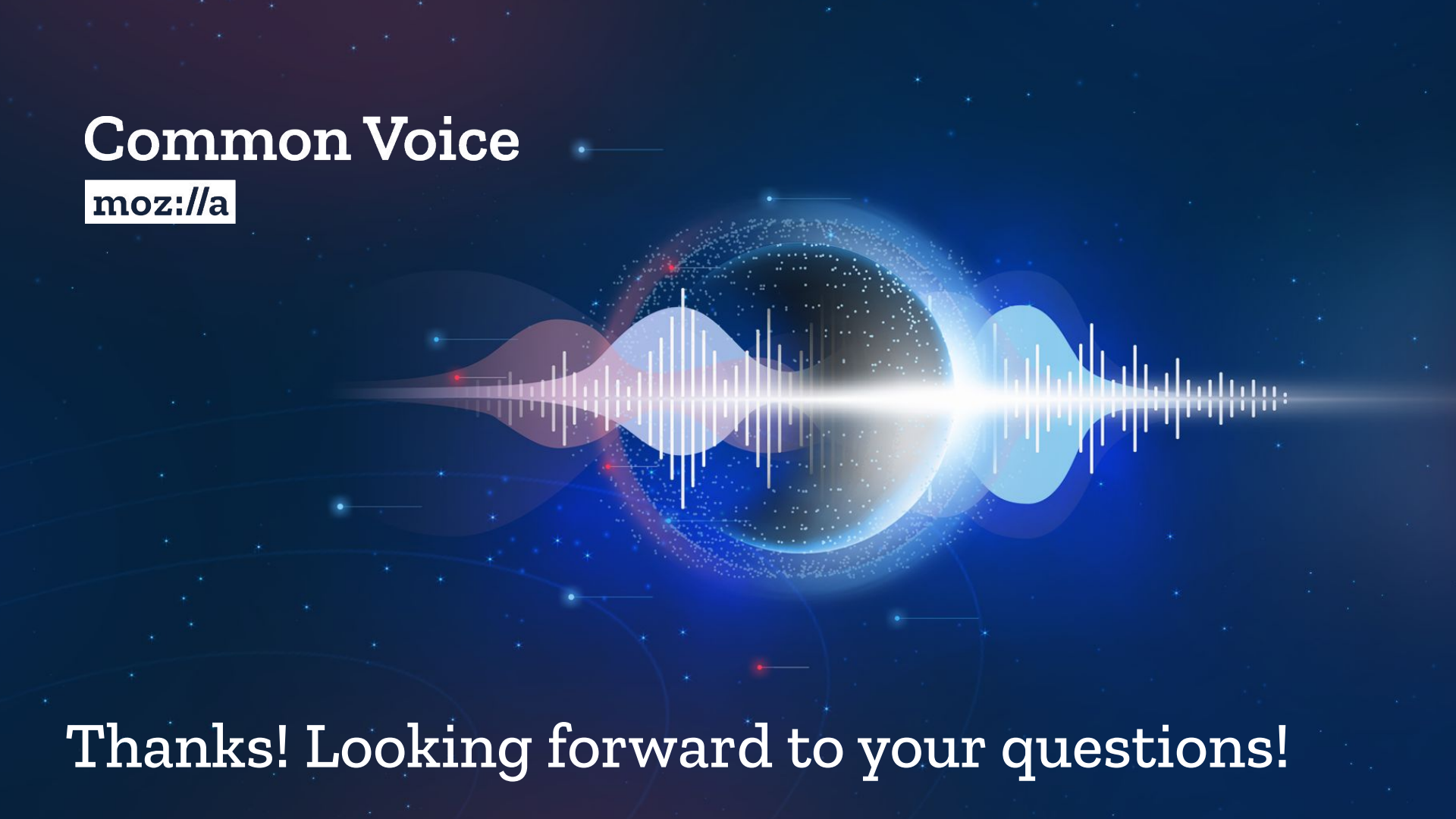
Contributing

Contribute Today

- Speak a language not covered by Sentence Extractor?
  - Help create a rules definition for that language
- Contribute PRs to Sentence Collector
  - Could benefit from some cleanup for the React part :)
- Help find good public domain sentences
- Help coordinate communities - Discourse forum
- Contribute your voice by recording sentences

Common Voice

moz://a



Thanks! Looking forward to your questions!