The Unsolvable Problem or the Unheard Answer?
A Dataset of 24,669 Open-Source Software Conference Talks

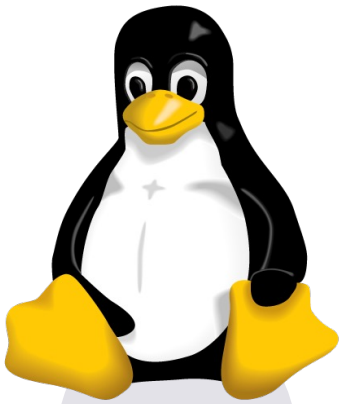
Kimberly Truong, Courtney Miller, Bogdan Vasilescu, Christian Kästner

Oregon State University

Carnegie Mellon University

What do we really know about open-source software?

- There's a direct correlation between community interactions and project success.
- There's a disconnect between practitioners and researchers
- What do practitioners experience? What do they want to talk about? What is **relevant**?
    - We don't know... and that's exactly what this dataset aims to minimize.



What do we provide?

Dataset

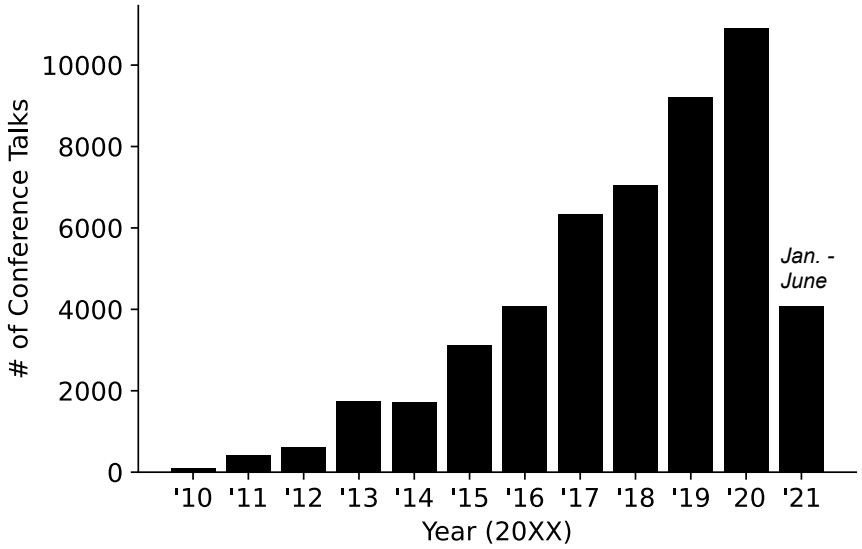
- Large and diverse
  - Disciplines
  - Size
  - Experiences/roles
  - Countries
  - Times
- Grey literature analysis
  - First-hand accounts
  - Captures what happened in time

Tool

- Extract YouTube video data
  - Video transcripts
  - Full videos
  - Expand beyond open-source software

Dataset

- 87 Conferences
- 24,669 Talks



Metadata

- Focus/theme
- Size
  - # of talks
  - # of speakers
  - # of attendees
- Affiliated conferences/organizations,
- Sponsorship information
- Main (or most recent)website
- Wikipedia page.

Methodology

1. Establishing a **conference list**
   - Google search for top 30 links
     - “open source conference”; “open source conference call for proposals”

2. Collecting metadata and **filtering** conferences:
   - Two documented editions
   - > 50 attendees or > 10 speakers/talks

3. **Compiling** data
   - Generate separate text files for

Resulting Data

- Name of video
- Publication date
- Playlist (often the conference edition)
- Description
- Transcript*
- YouTube URL

* Not shared with the Dataset due to the YouTube License

```
Title: Russell Keith-Magee - Keynote - PyCon 2019
Publication date: 2019-05-05
Playlist: PyCon 2019 - Keynotes
Description:
        "Speaker: Russell Keith-Magee

Keynote

Slides can be found at: https://speakerdeck.com/pycon2019 and
Captions:
        00:00:41,270 --> 00:00:51,399
good morning

00:00:42,940 --> 00:00:53,800
[Applause]

00:00:51,399 --> 00:01:00,820
I couldn't be more excited to welcome

00:00:53,800 --> 00:01:03,699
you to the 2019 Pike on pike on 2019

00:01:00,820 --> 00:01:05,770
here in Cleveland Ohio on behalf of the

00:01:03,699 --> 00:01:08,380
pike on 2019 staff I want to start by

00:01:05,770 --> 00:01:09,610
saying thank you thank you to all the

00:01:08,380 --> 00:01:11,770
volunteers that make this conference

00:01:09,610 --> 00:01:13,509
possible to the Python Software

00:01:11,770 --> 00:01:15,820
Foundation for taking on the fiscal
```

Possible Applications

- Analyzing popularity and diffusion
- Identifying major influences on communities
- Understanding common challenges
- Identifying trends over time
- Filtering to create samples by time, focus, size, etc.

How did we use it?

- Topic modeling
- Generating samples from the larger set
  - Can be expanded with topic modeling
  - Sample of practitioners sharing why they left open-source software
- See the Disengagement Diaries here

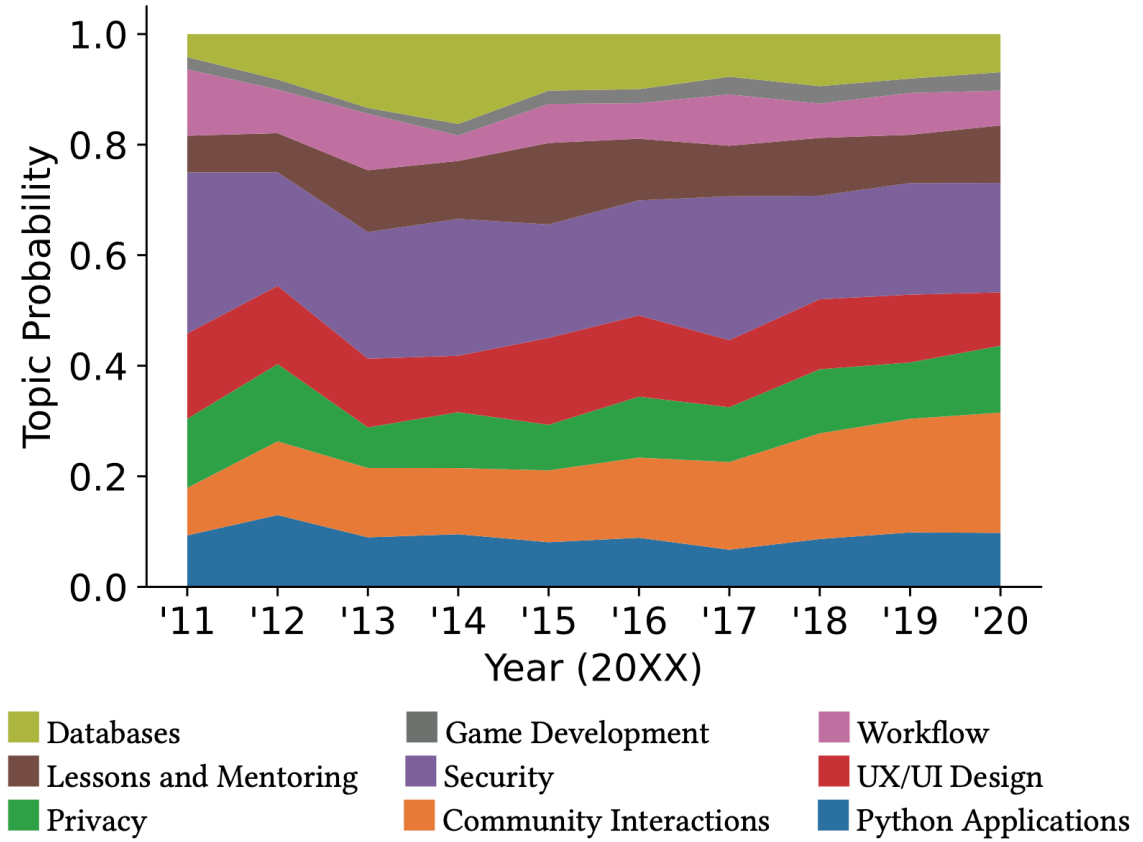
Topic Modeling

Method

1. Preprocess text
2. Trial and error with 7 to 40 topics.
3. Consolidate topics

Observations

- Community interactions is a more common topic
- Talks about Databases have decreased



Thank you!

e. truonkim@oregonstate.edu

Carnegie
Mellon
University