

# The Unsolvable Problem or the Unheard Answer?

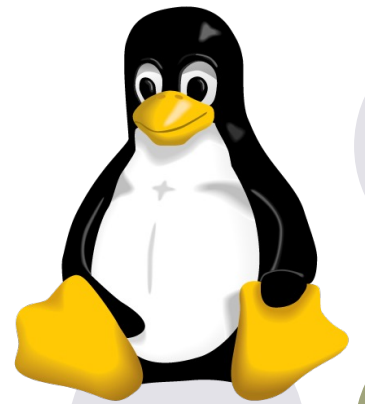
A Dataset of 24,669 Open-Source Software Conference Talks

Kimberly Truong, Courtney Miller, Bogdan Vasilescu, Christian Kästner



# What do we really know about open-source software?

- 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.



# Past Works

- Study common issues, what leads to success, etc.
- Limited in scope and sample size

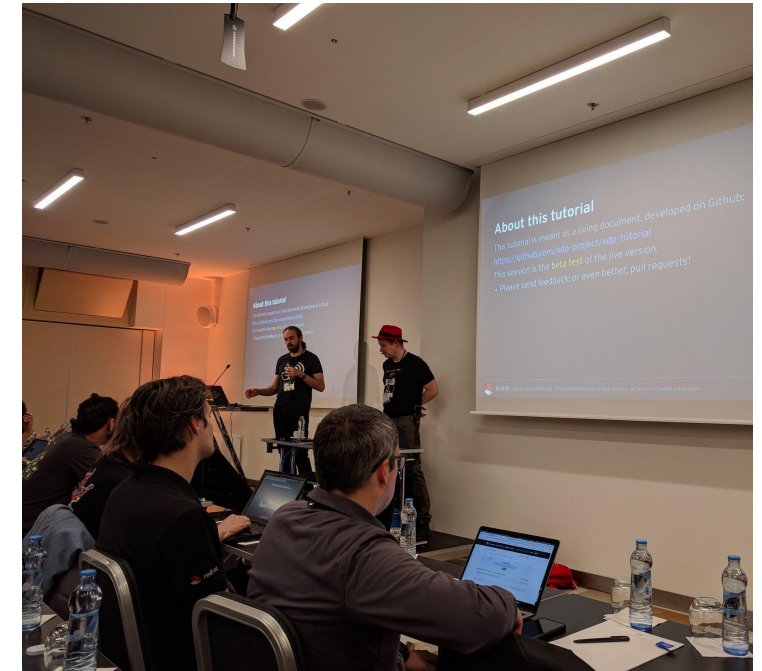


PyCon Africa 2019,  
<https://www.ict4d.at/tag/pycon/>

3



Open Source Summit 2019,  
[https://live.staticflickr.com/65535/51647514916\\_acc4a7e90b\\_c.jpg](https://live.staticflickr.com/65535/51647514916_acc4a7e90b_c.jpg)



NetDev 2019,  
<https://bootlin.com/blog/tag/netdev/>

# What do we provide?

## Dataset

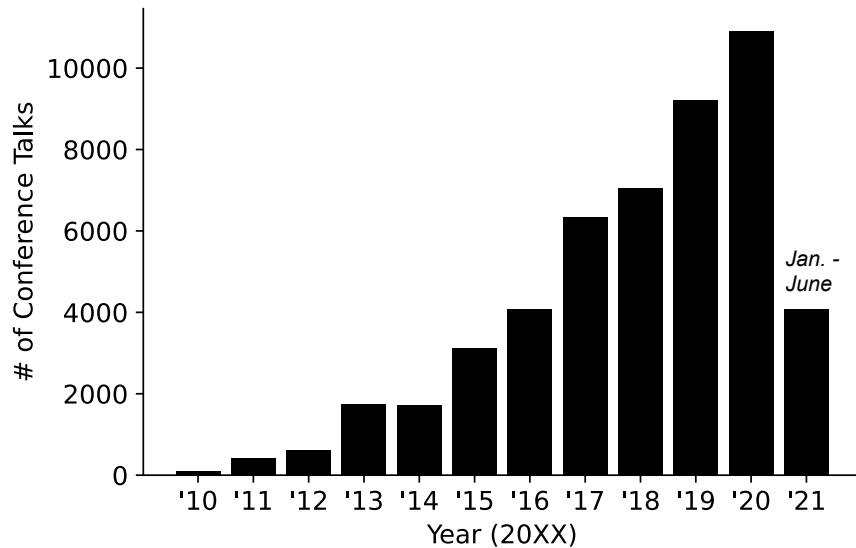
- Large and diverse
  - Disciplines
  - Size
  - Experiences/roles
  - Countries
  - Times
- Grey literature analysis
  - First-hand accounts
  - Captures what happened in time

## 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 each video.

# 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

```
1 Title: I [REDACTED] Keynote - PyCon 2019
2 Publication date: 2019-05-05
3 Playlist: PyCon 2019 - Keynotes
4 Description:
5     "Speaker: [REDACTED]"
6
7 Keynote
8
9 Slides can be found at: https://speakerdeck.com/pycon2019 and https://
10 Captions:
11     00:00:41,270 --> 00:00:51,399
12     good morning
13
14     00:00:42,940 --> 00:00:53,800
15     [Applause]
16
17     00:00:51,399 --> 00:01:00,820
18     I couldn't be more excited to welcome
19
20     00:00:53,800 --> 00:01:03,699
21     you to the 2019 Pike on pike on 2019
22
23     00:01:00,820 --> 00:01:05,770
24     here in Cleveland Ohio on behalf of the
25
26     00:01:03,699 --> 00:01:08,380
27     pike on 2019 staff I want to start by
28
29     00:01:05,770 --> 00:01:09,610
30     saying thank you thank you to all the
31
32     00:01:08,380 --> 00:01:11,770
33     volunteers that make this conference
34
35     00:01:09,610 --> 00:01:13,509
36     possible to the Python Software
37
38     00:01:11,770 --> 00:01:15,820
39     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.



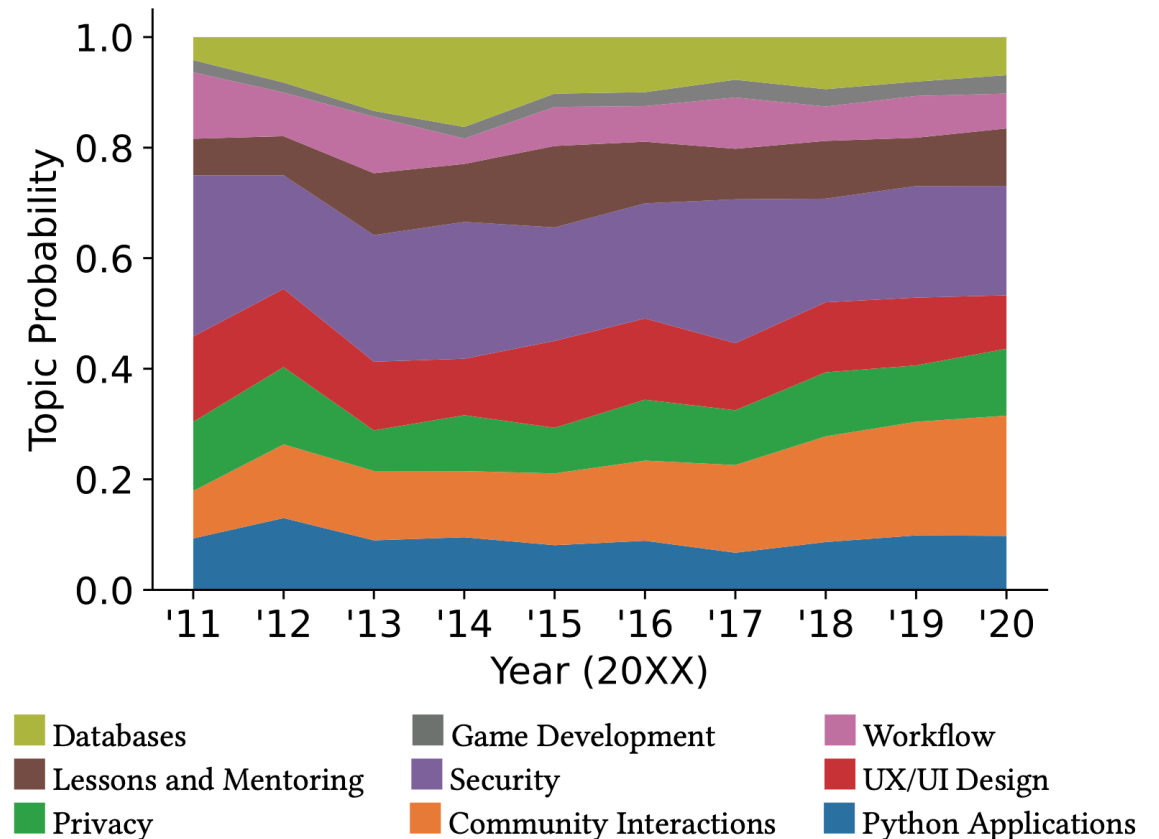
# Topic Modeling

## 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



# Disengagement Diaries

## 3 Major Factors:

- External
- Cultural
- Volunteering



<https://disengagement-diaries.github.io>

Not Enough Time (Internal) No Longer Enjoyable

## Who?

**GitHub:** Removed for privacy

**Project:** [Django](#)

## Why?

Anna cited three reasons for leaving their open source communities at DjangoCon US in 2018. These reasons arose because they had to change their priorities as new events unfolded in their life. The first issue was the lack of a work life balance as newer, more important responsibilities arose.

*"People's living situations change or your priorities change and so people may say after a while I can't do it anymore for whatever reason." (15:44)*

*"Tech kind of takes over our whole life. I'm definitely guilty of that so I'm kind of like actively saying no." (30:00)*

This led to a lack of time internally due to their shift in priorities. Although they still appreciated the Django community, Anna found contributing to open source to take away from more significant priorities, like a new job and changing living situations.

*"I'm actually moving cities so I also kind of want to prioritize my personal life a little more I've served the community a lot and now it's kind of like time to look after myself a little more I'm out I also took on this*

# Summary



Dataset of 24,669 open-source software conference talks



New source of information for grey literature analysis



Bridge the gap between open-source software research and practice

e. [truonkim@oregonstate.edu](mailto:truonkim@oregonstate.edu)

GitHub: /KimberlyTruong

LinkedIn: /in/kimberly-le-truong-cs

