The Unsolvable Problem or the Unheard Answer?

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

Kimberly Truong,[1] Courtney Miller,[2] Bogdan Vasilescu,[2] Christian Kästner [2]

[1] Oregon State University, USA [2] Carnegie Mellon University, USA

Oregon State University

Carnegie Mellon University

Motivations

There's a disconnect between open-source researchers and practitioners.

- There is no correlation between the perceived relevance of a software engineering conference paper and its number of citations.
- Questions relevant to practitioners were rarely found in research papers.

Community interactions in a project are an indication of the project's success.

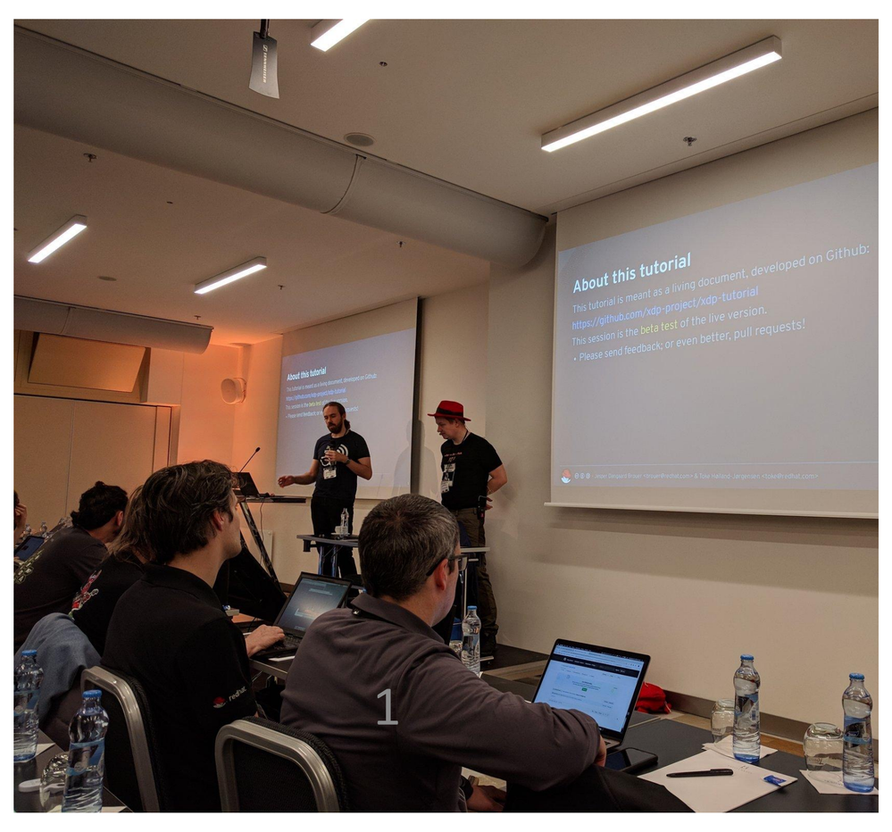
We want to better understand practitioners through what they share.

Source of Grey literature

- Large and diverse untapped source(24,669 talks) publicly available online
  - Disciplines – different tools, specialties, fields
  - Size – small (50 attendees) to large (~ 4,000 attendees)
  - Experiences/roles – contributor, core maintainer, user, CEO
  - Regions – North America, Europe, Australia, parts of Africa
  - Times – 10 years (2010 - 2021)
- Provides first-hand accounts from practitioners

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

Open Source Summit 2019, https://live.staticflickr.com/65535/51647514916_acc4a7e90b_c.jpg



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

Dataset

Information

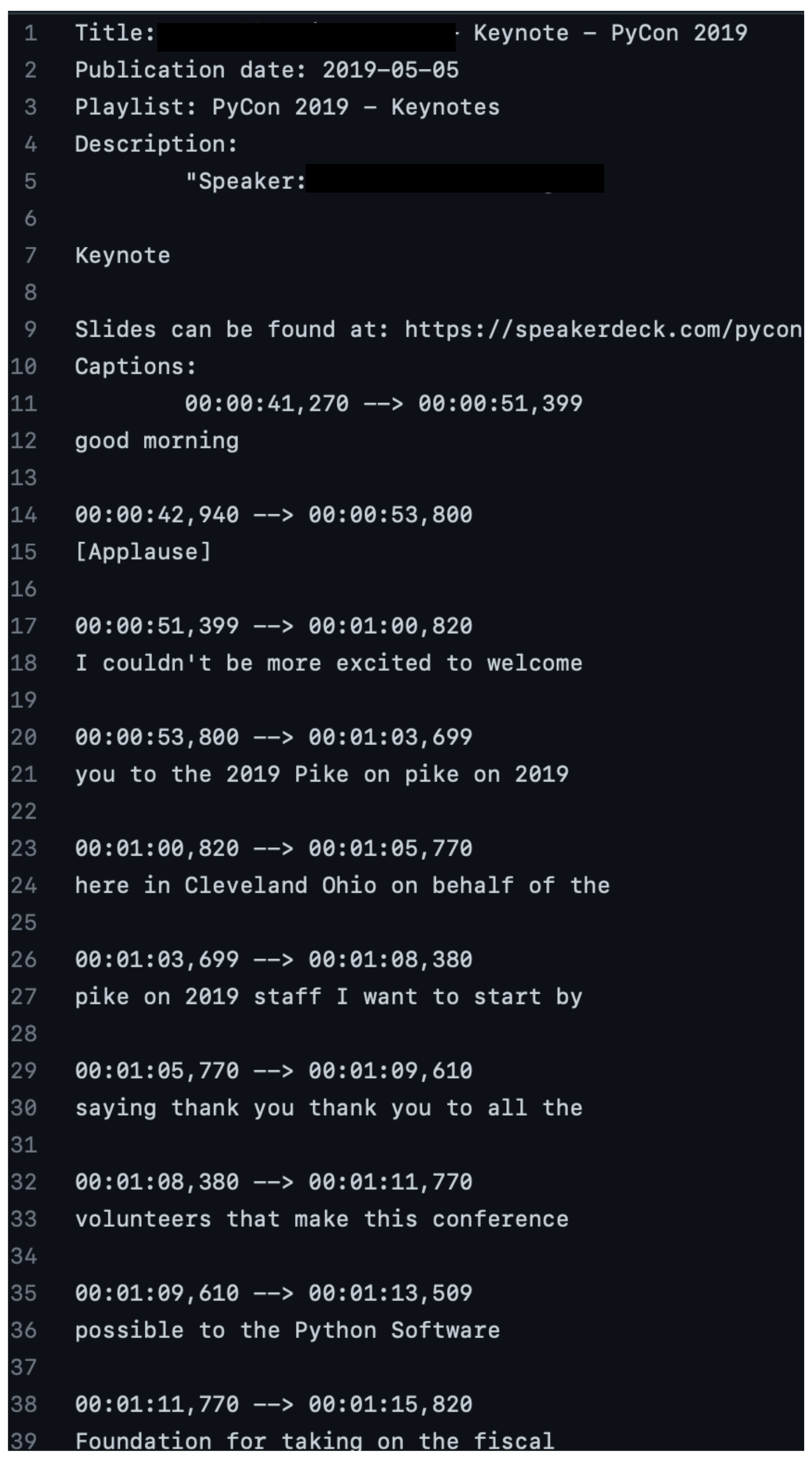
- Name of video
- Publication date
- Playlist (conference edition)
- Description
- Transcript*
- YouTube URL

Metadata

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

Tool

- Scripts to extract YouTube video data

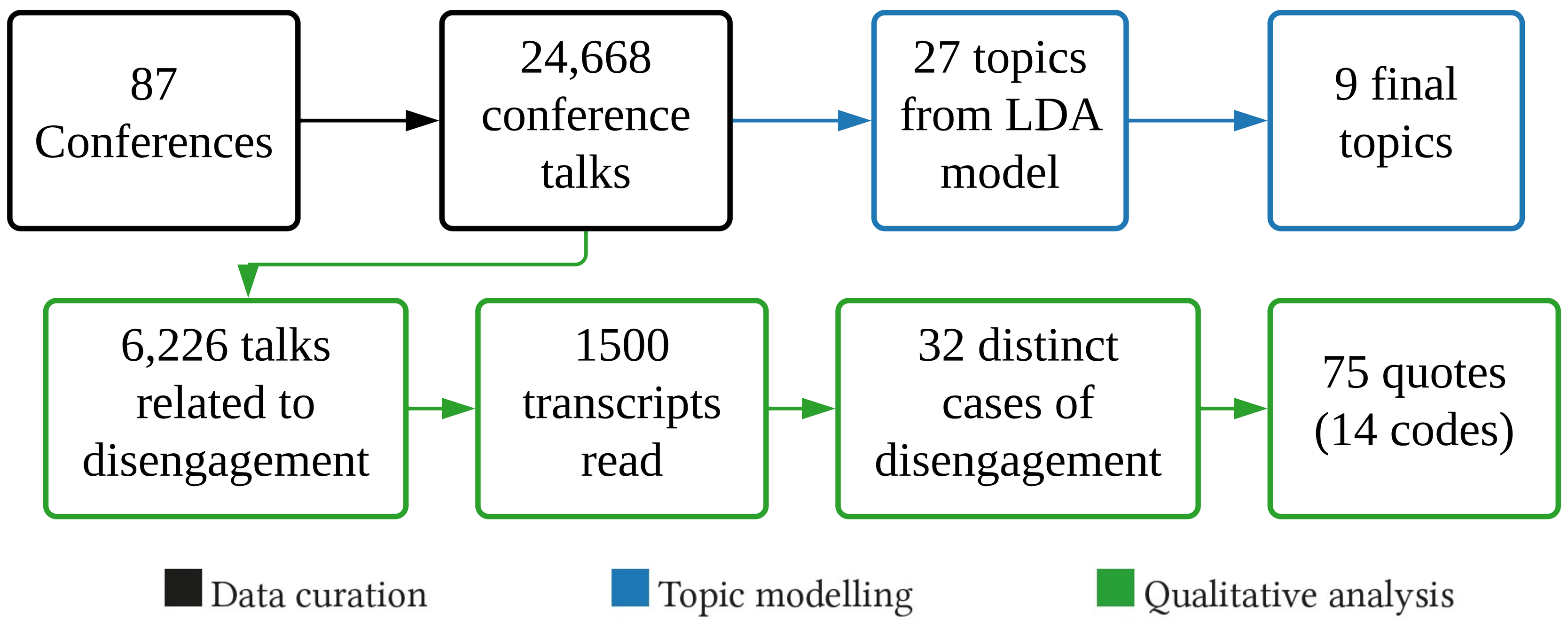


```
Title: [redacted] Keynote – PyCon 2019
Publication date: 2019-05-05
Playlist: PyCon 2019 – Keynotes
Description:
        "Speaker: [redacted]

Keynote

Slides can be found at: https://speakerdeck.com/pycon
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
```

* Not shared with the Dataset due to the YouTube License

Methodology

1. Establish a **conference list**
   - Google search for top 30 links
     - "open source conference"; "open source conference call for proposals"
2. Collect metadata and **filtering** conferences:
   - Two documented editions
   - 50 attendees or > 10 speakers/talks
3. **Compile** data
   - Generate separate text files for each video



Applications

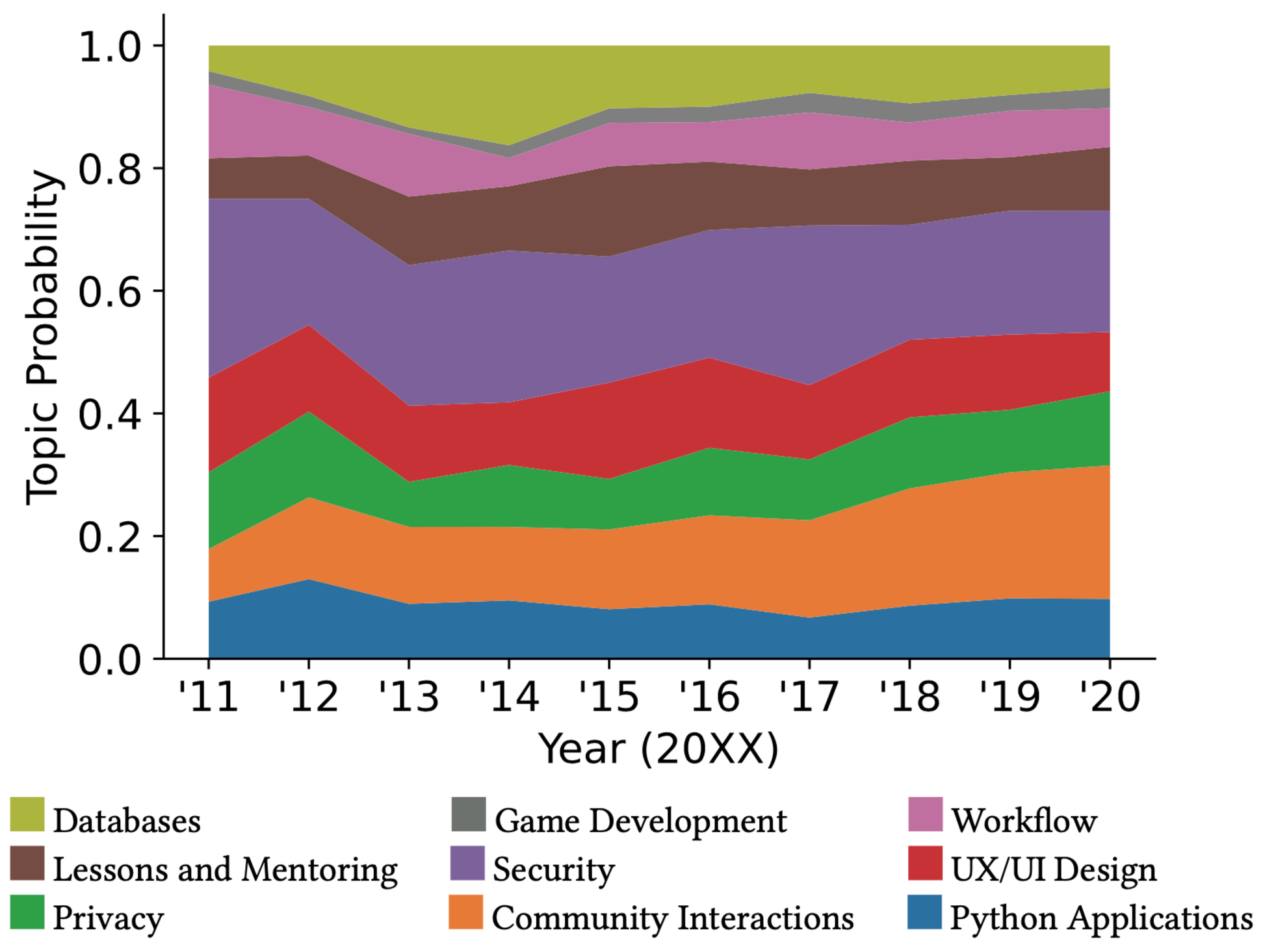
- Analyze the popularity and diffusion of tools and practices
- Understand common challenges discussed by practitioners
- Identify trends over time
- Identify major influences on certain communities
  - e.g., how the conference impacted the contributions for a certain project
- Filter to create samples by time, focus, size, etc.

Topic Model

We see how the frequency of the most common topics in open-source talks have changed over 10 years.

Observations

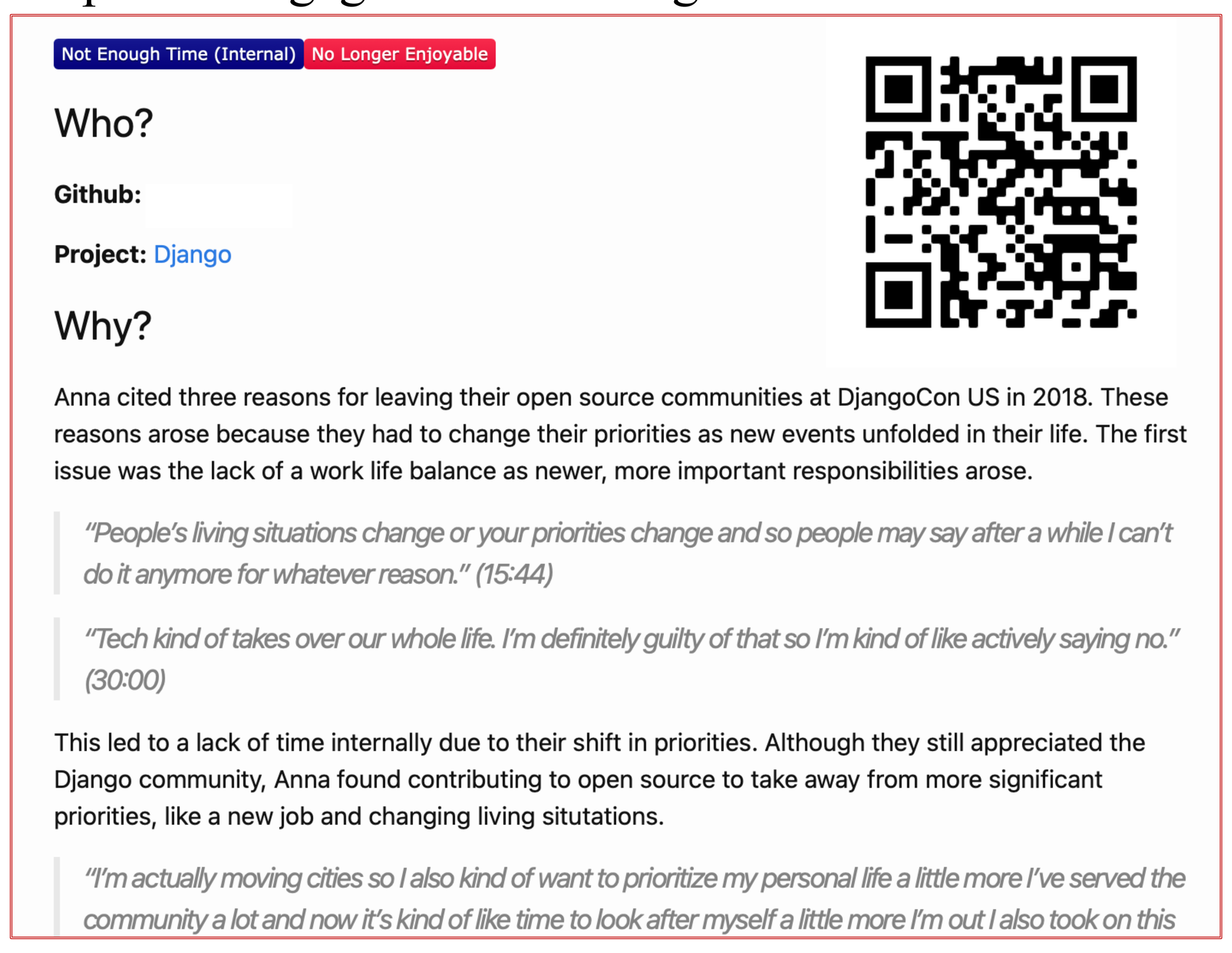
- Community interactions is increasingly common.
- Talks about Databases have decreased after 2014.



Disengagement Diaries

1. Generate a sample
   - Keywords: 'leave', 'abandon', 'hostility', etc.
2. Thematically code relevant transcripts.
3. Record reasons cited for disengagement, GitHub activity, and interventions

https://disengagement-diaries.github.io



Cited Factors

- Cultural (32%)
  - Community leadership, policy disagreements, project direction, hostility
- External (18%)
  - Not enough time, self-doubt, health, left company, no longer useful
- Volunteering (50%)
  - Not enough time (internal), lack of support, no longer enjoyable, guilt, burnout

For more information…

Read our paper at https://doi.org/10.1145/3524842.3528488

e. truonkim@oregonstate.edu /kimberly-le-truong-cs /KimberlyTruong