AI | BLINDSPOT

A discovery process for spotting unconscious biases and structural inequalities in AI systems

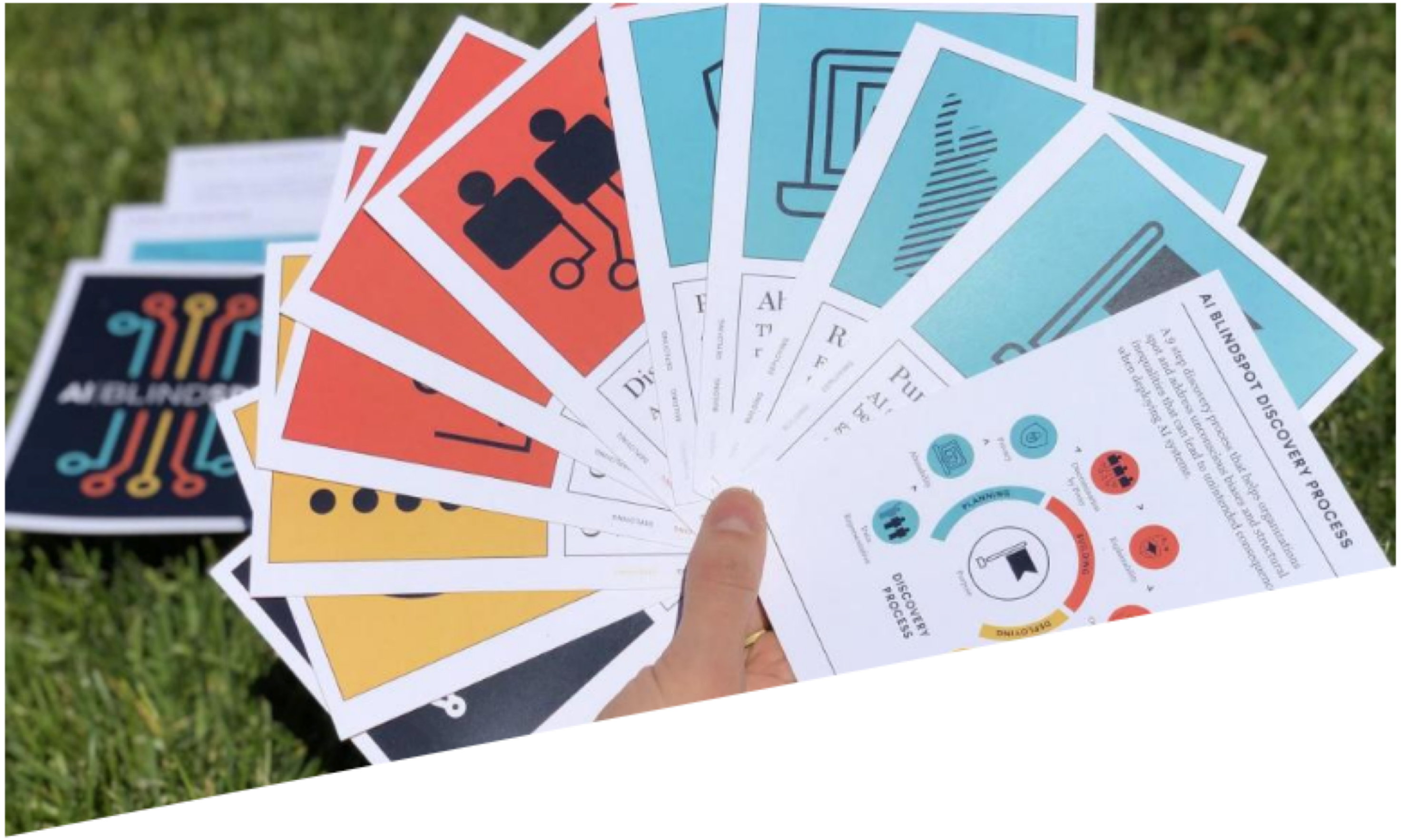
Hong Qu
CS 105 Harvard University

October 22, 2019

Overview

- Introduction
- AI Blindspot project
    - Spotting Risk in the machine learning
- Case study: YouTube video recommendations
- Group exercise

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

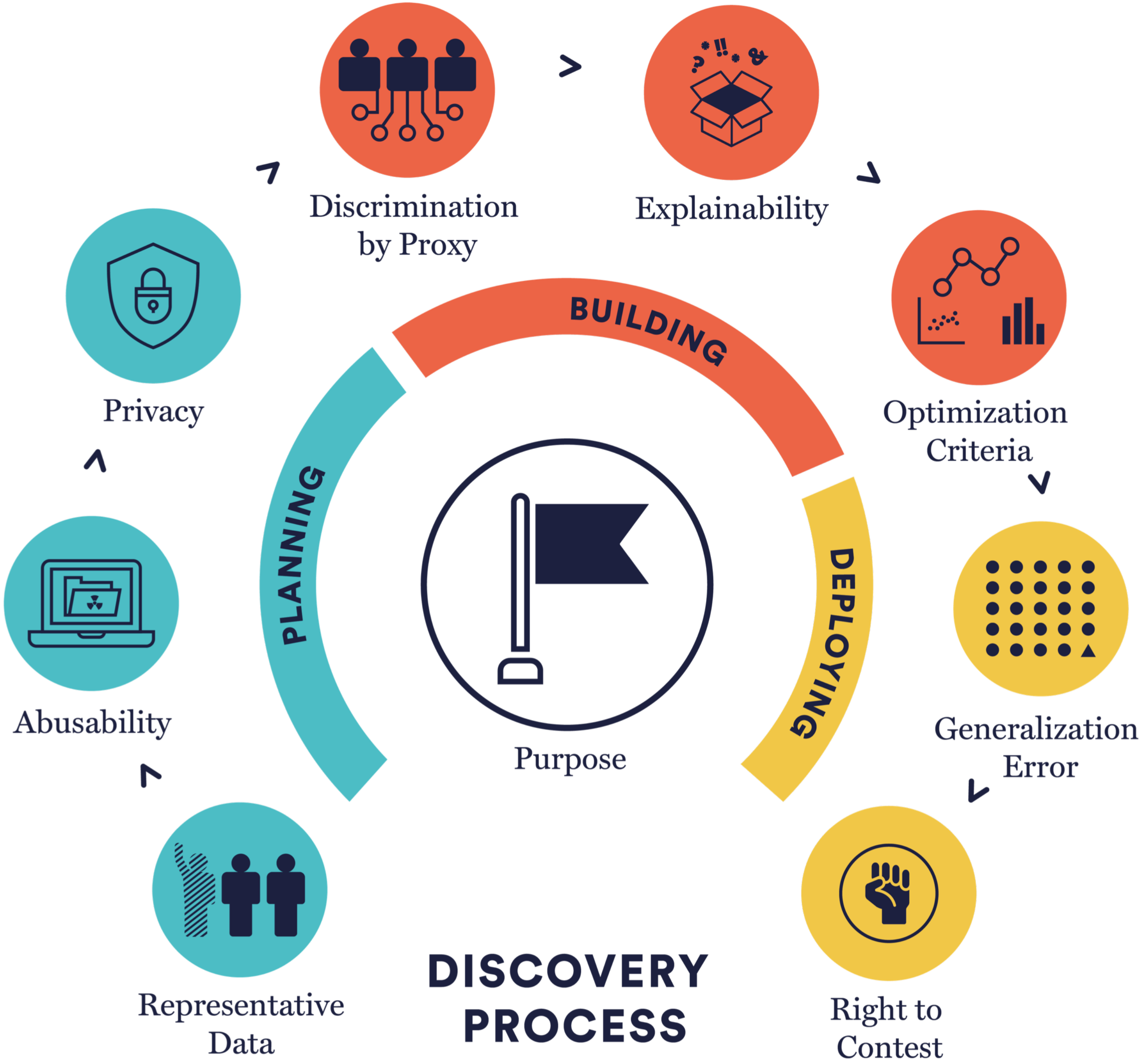

AI BLINDSPOT
AI BLINDSPOT DISCOVERY PROCESS
A 9 step discovery process that helps organizations spot and address unconscious biases and structural inequalities that can lead to unintended consequences when deploying AI systems.
PLANNING
BUILDING
DEPLOYING
DISCOVERY PROCESS

HOME › NEWS ›

What We Created During Assembly 2019

Blindspot Discovery Process

1. Purpose
2. Representative Data
3. Abusability
4. Privacy
5. Discrimination by Proxy
6. Explainability
7. Optimization Criteria
8. Generalization Error
9. Right to Contest



Root causes of algorithmic bias

Unconscious bias

- Diversity of teams
- Proxy variables
- Real world context vs ideal models
- Lack of control over AI/ML pipeline
- Tradeoffs between competing goals
- Etc…

Structural inequalities

- Historical legacy in society
- Civil rights and anti-discrimination
- Protected classes
- Inclusion and equity
- Shifts in values
- Etc…

Case Study: YouTube recommendation algorithm

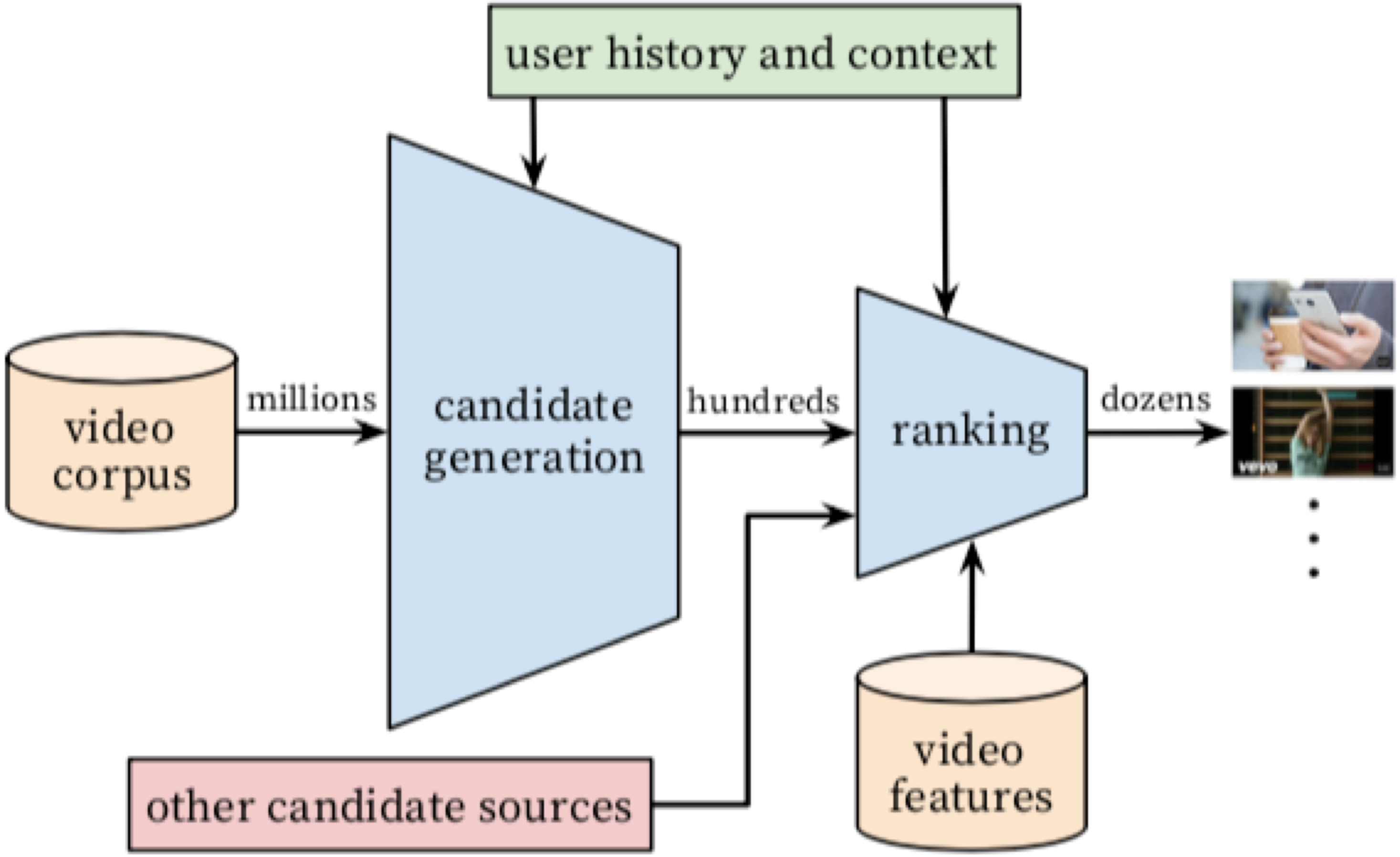
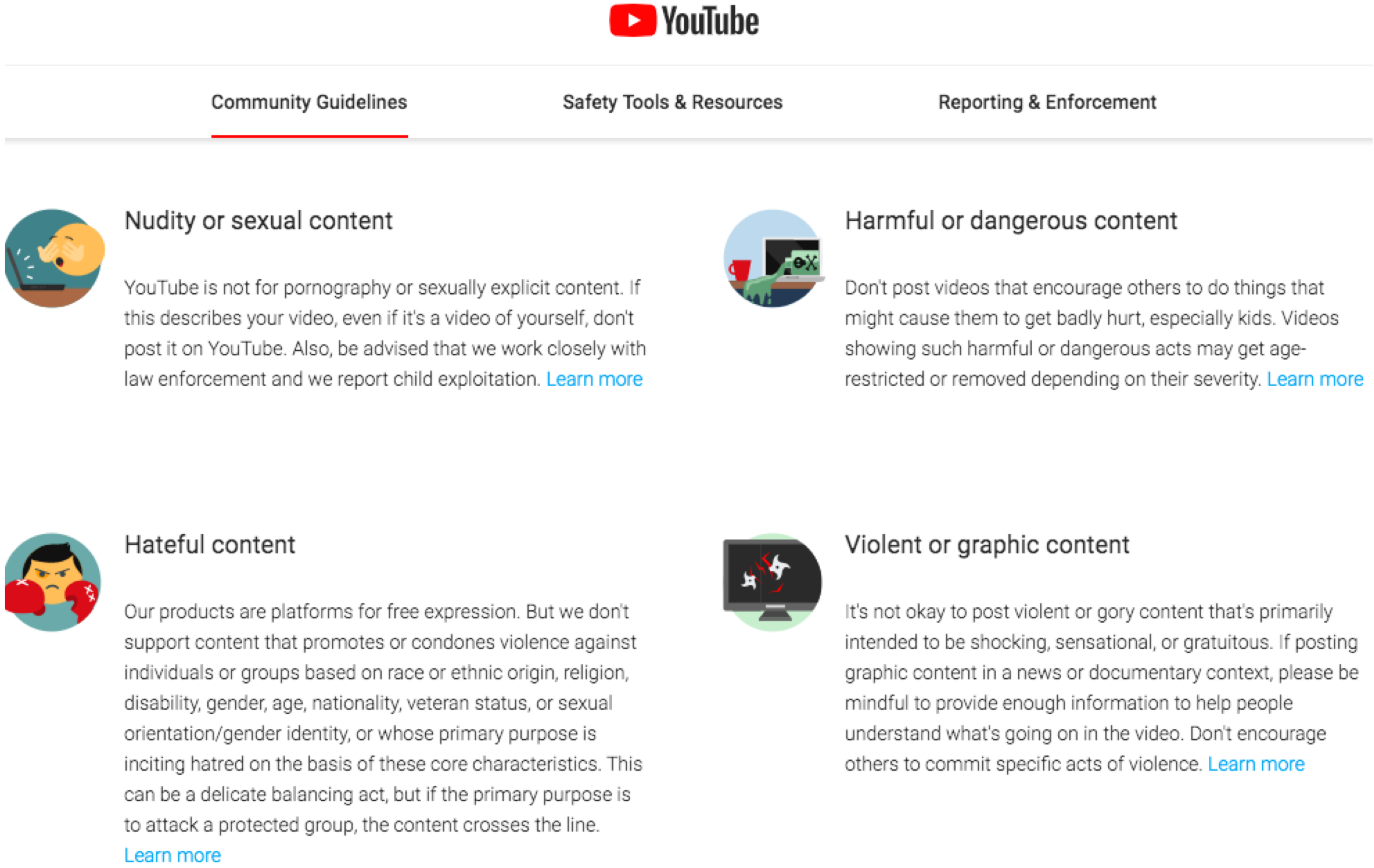


Figure 2: Recommendation system architecture demonstrating the "funnel" where candidate videos are retrieved and ranked before presenting only a few to the user.



Nudity or sexual content

YouTube is not for pornography or sexually explicit content. If this describes your video, even if it's a video of yourself, don't post it on YouTube. Also, be advised that we work closely with law enforcement and we report child exploitation. Learn more

Harmful or dangerous content

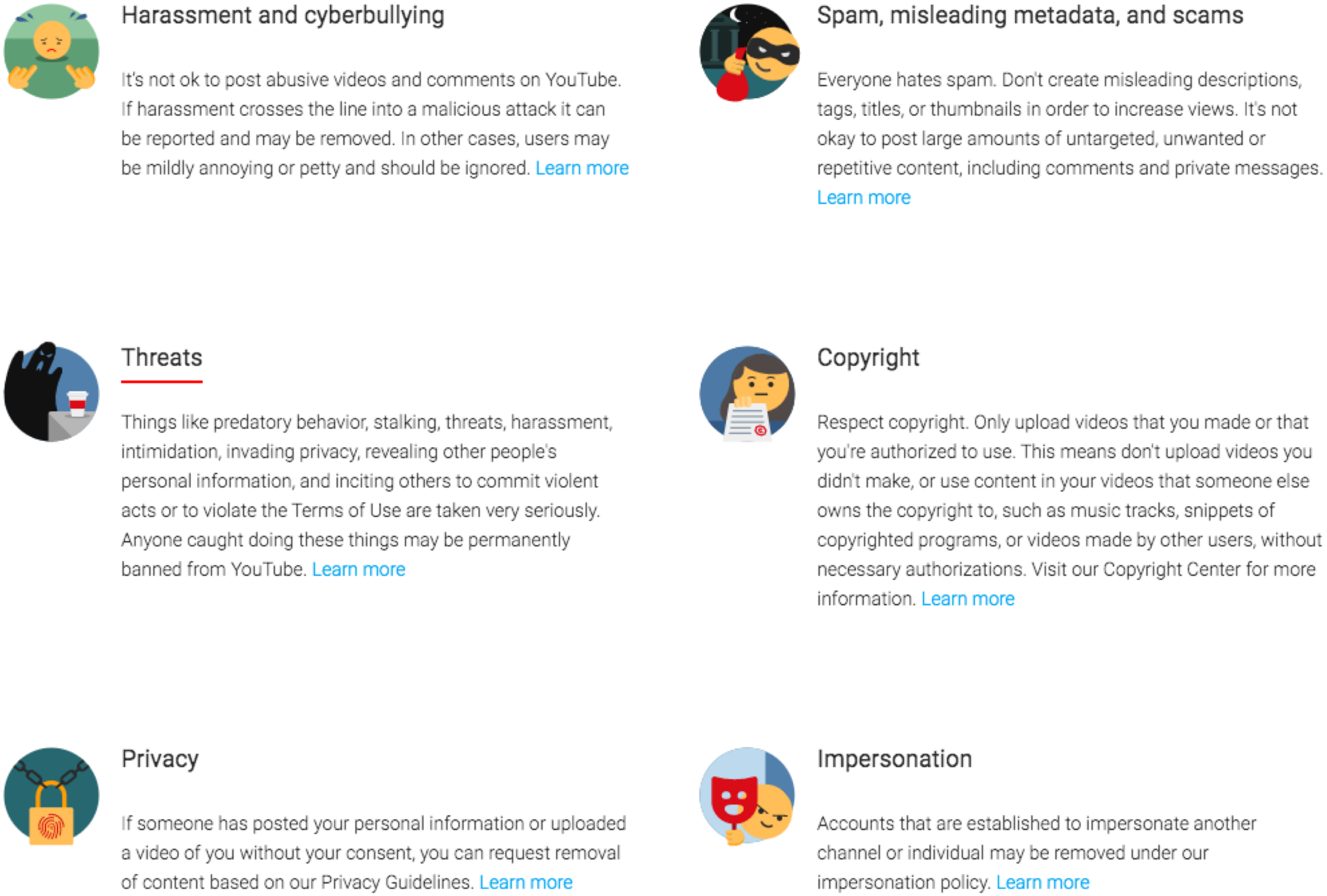
Don't post videos that encourage others to do things that might cause them to get badly hurt, especially kids. Videos showing such harmful or dangerous acts may get age-restricted or removed depending on their severity. Learn more

Hateful content

Our products are platforms for free expression. But we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics. This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line. Learn more

Violent or graphic content

It's not okay to post violent or gory content that's primarily intended to be shocking, sensational, or gratuitous. If posting graphic content in a news or documentary context, please be mindful to provide enough information to help people understand what's going on in the video. Don't encourage others to commit specific acts of violence. Learn more



Harassment and cyberbullying

It's not ok to post abusive videos and comments on YouTube. If harassment crosses the line into a malicious attack it can be reported and may be removed. In other cases, users may be mildly annoying or petty and should be ignored. Learn more

Threats

Things like predatory behavior, stalking, threats, harassment, intimidation, invading privacy, revealing other people's personal information, and inciting others to commit violent acts or to violate the Terms of Use are taken very seriously. Anyone caught doing these things may be permanently banned from YouTube. Learn more

Privacy

If someone has posted your personal information or uploaded a video of you without your consent, you can request removal of content based on our Privacy Guidelines. Learn more

Spam, misleading metadata, and scams

Everyone hates spam. Don't create misleading descriptions, tags, titles, or thumbnails in order to increase views. It's not okay to post large amounts of untargeted, unwanted or repetitive content, including comments and private messages. Learn more

Copyright

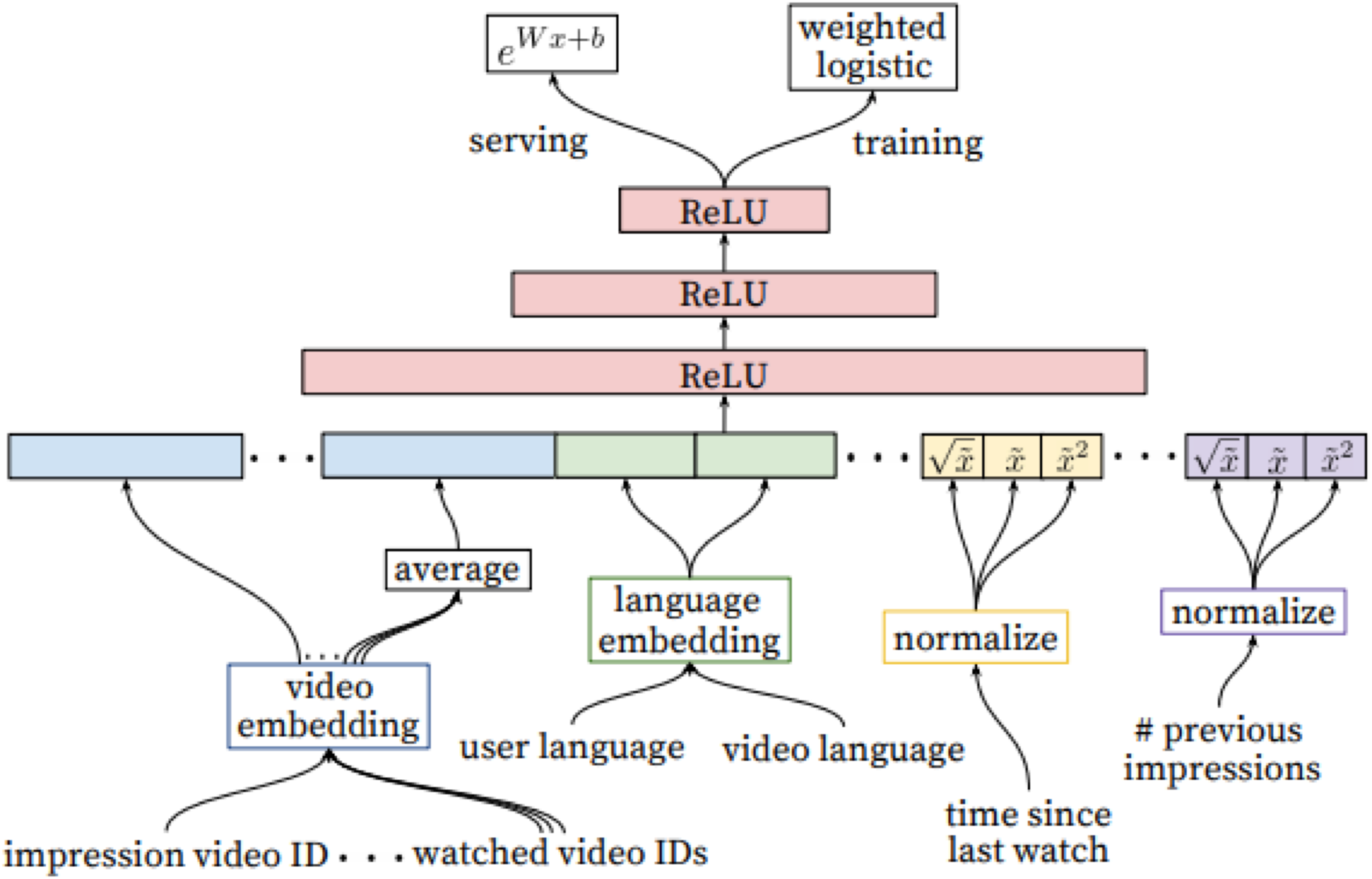
Respect copyright. Only upload videos that you made or that you're authorized to use. This means don't upload videos you didn't make, or use content in your videos that someone else owns the copyright to, such as music tracks, snippets of copyrighted programs, or videos made by other users, without necessary authorizations. Visit our Copyright Center for more information. Learn more

Impersonation

Accounts that are established to impersonate another channel or individual may be removed under our impersonation policy. Learn more

Deep Neural Networks for YouTube Recommendations

Paul Covington, Jay Adams, Emre Sargin
Google



https://ai.google/research/pubs/pub45530

Feature selection

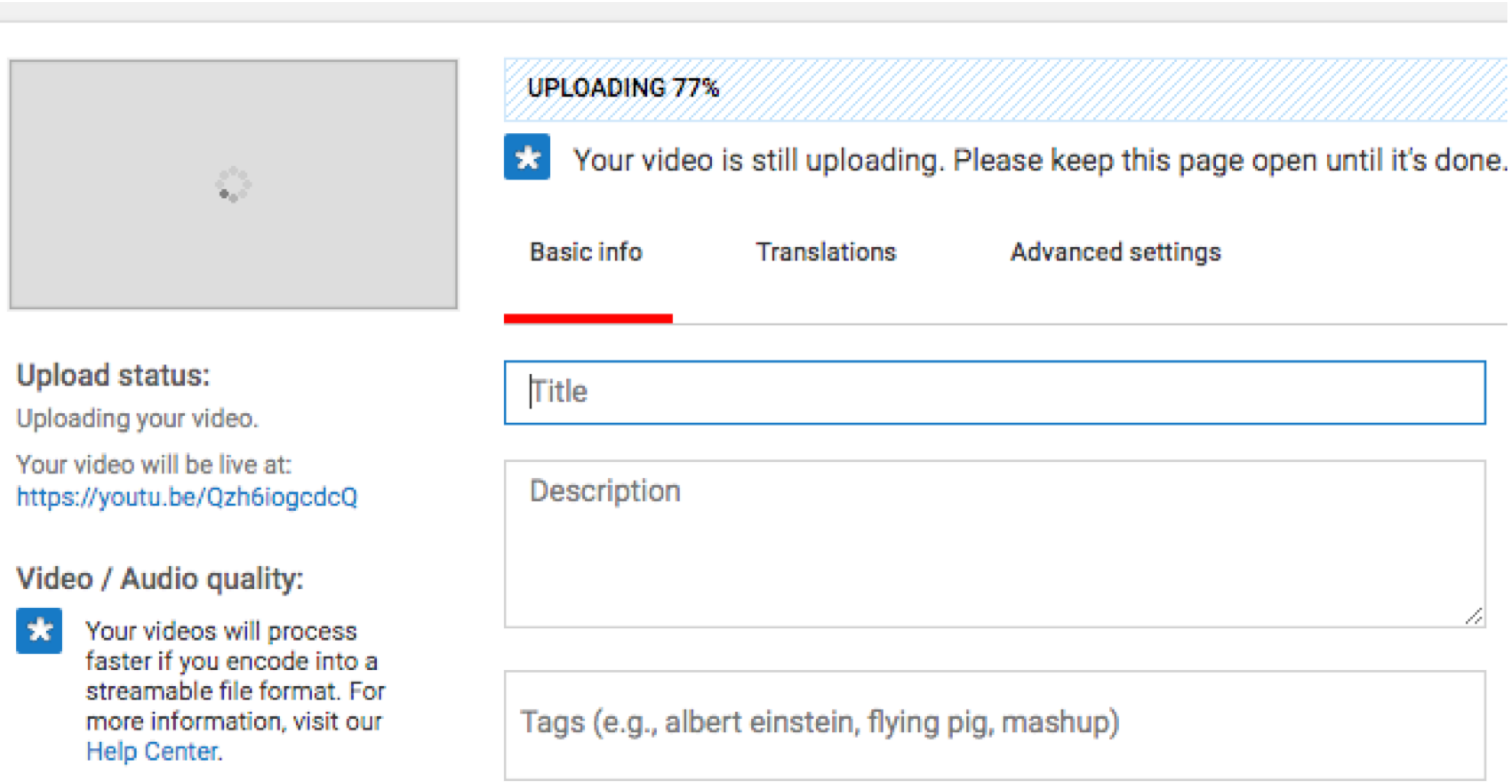
Content signals

- Metadata
- Collaborative filtering
- Embedded video views
- Freshness
- Authoritative sources
- Etc…

User signals

- Video view history
- Search history
- Location and language
- Time since last watched
- Age
- Etc…

Abusability



UPLOADING 77%

Your video is still uploading. Please keep this page open until it's done.

Title

Description

Tags (e.g., albert einstein, flying pig, mashup)

Upload status:

Uploading your video.

Your video will be live at:
https://youtu.be/Qzh6iogcdcQ

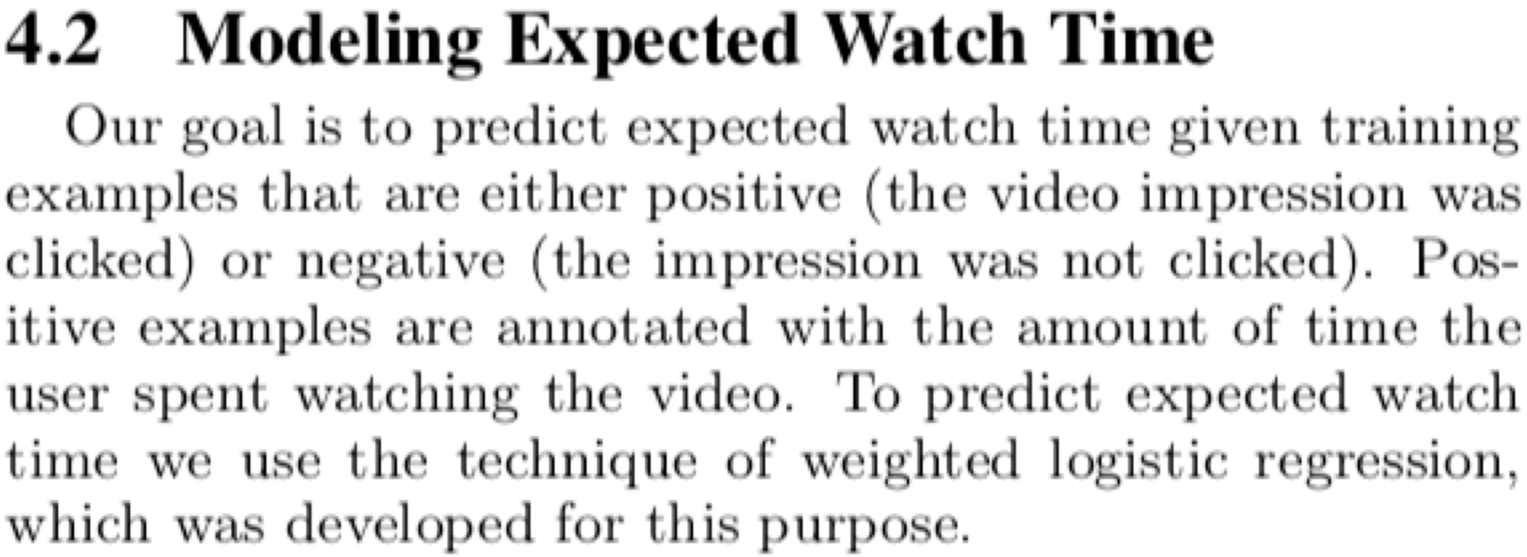
Video / Audio quality:

Your videos will process faster if you encode into a streamable file format. For more information, visit our Help Center.

Ranking by **click-through rate often promotes deceptive videos** that the user does not complete (“clickbait”) whereas watch time better captures engagement

Optimization Criteria

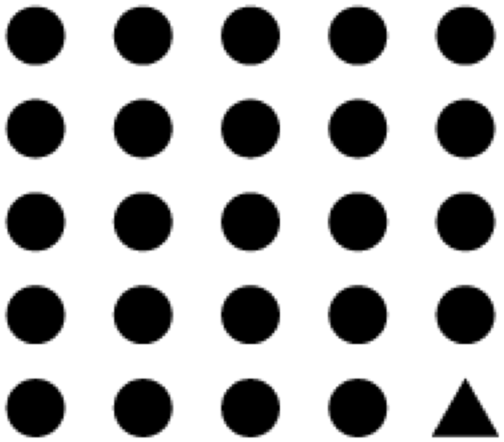
Ranking by click-through rate often promotes deceptive videos that the user does not complete (“clickbait”) whereas watch time better captures engagement

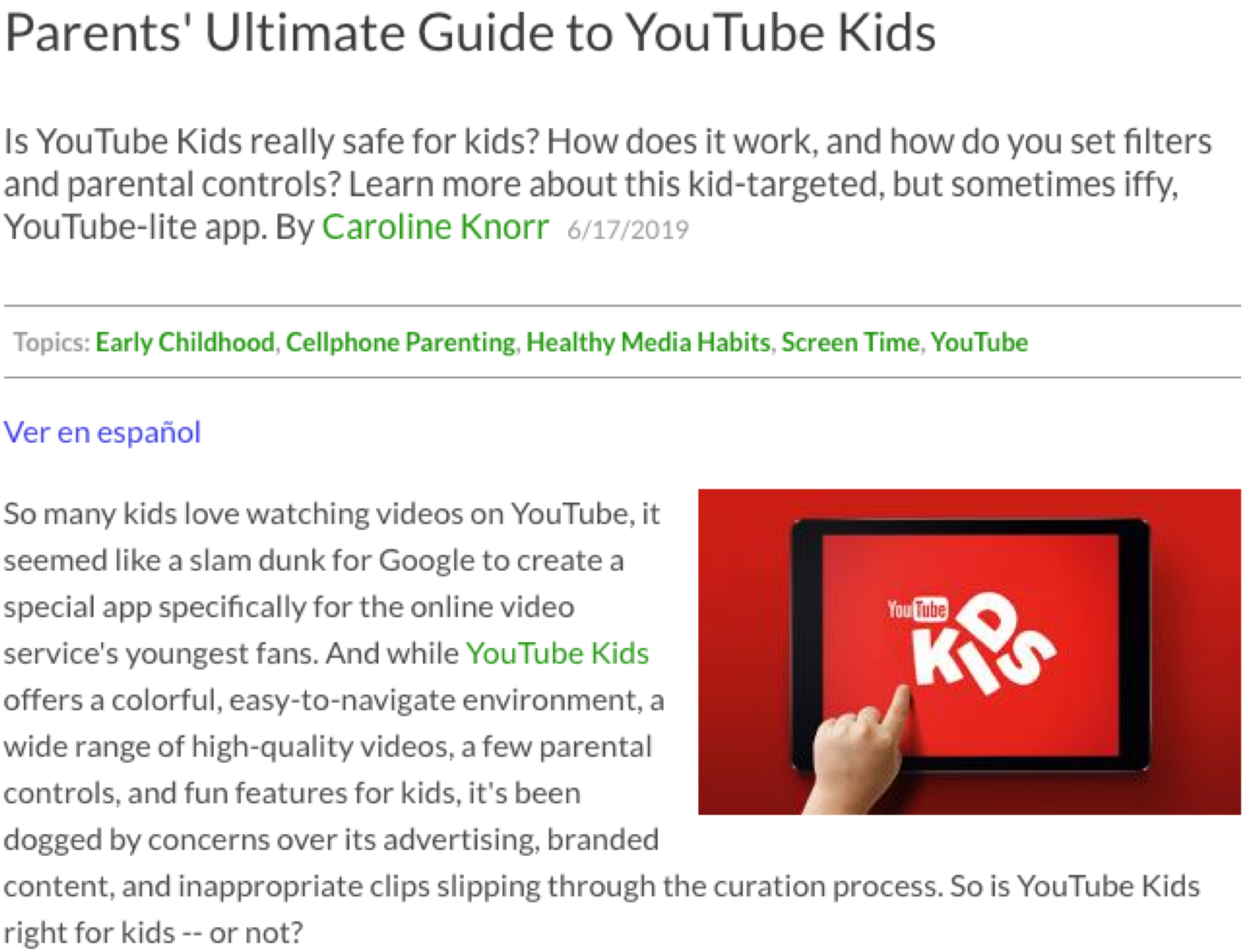


4.2 Modeling Expected Watch Time

Our goal is to predict expected watch time given training examples that are either positive (the video impression was clicked) or negative (the impression was not clicked). Positive examples are annotated with the amount of time the user spent watching the video. To predict expected watch time we use the technique of weighted logistic regression, which was developed for this purpose.

Generalization Error





Parents' Ultimate Guide to YouTube Kids

Is YouTube Kids really safe for kids? How does it work, and how do you set filters and parental controls? Learn more about this kid-targeted, but sometimes iffy, YouTube-lite app. By Caroline Knorr 6/17/2019

Topics: Early Childhood, Cellphone Parenting, Healthy Media Habits, Screen Time, YouTube

Ver en español

So many kids love watching videos on YouTube, it seemed like a slam dunk for Google to create a special app specifically for the online video service's youngest fans. And while YouTube Kids offers a colorful, easy-to-navigate environment, a wide range of high-quality videos, a few parental controls, and fun features for kids, it's been dogged by concerns over its advertising, branded content, and inappropriate clips slipping through the curation process. So is YouTube Kids right for kids -- or not?

Purpose

What is the purpose of the YouTube recommendation algorithm?

???

Group Exercise

- Form groups of 5 people
- Choose an application of machine learning in the following domains
    - Employment
    - Education
    - Finance
    - Healthcare
    - Courts
    - Housing
    - Facial recognition
- Use the AI Blindspot cards to spot and mitigate risks in the AI system