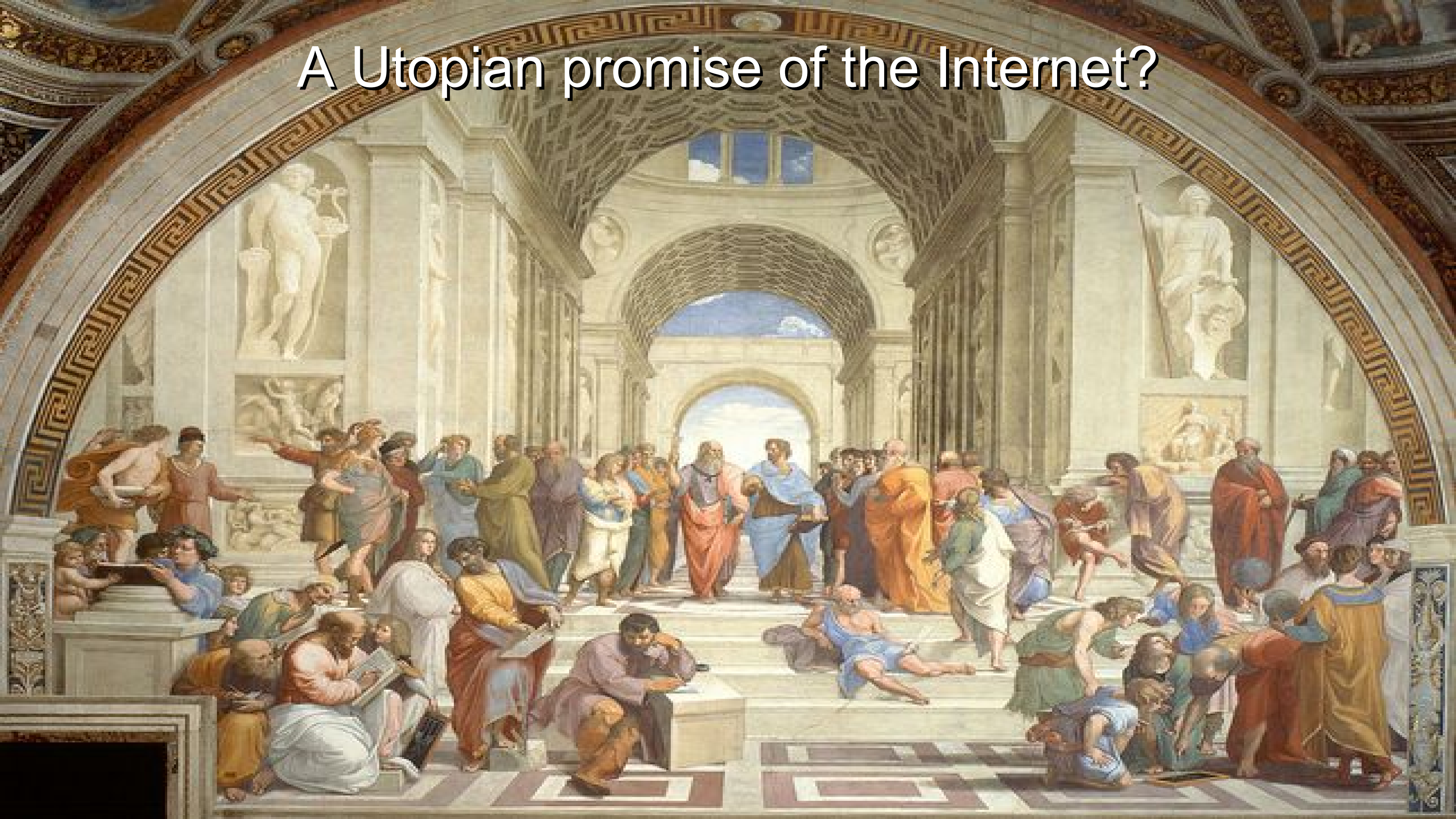
Conversation Corpora, Emotional Robots, and Battles with Bias

Lucas Dixon @Wikimedia, Nov 2017

ConversationAI.github.io

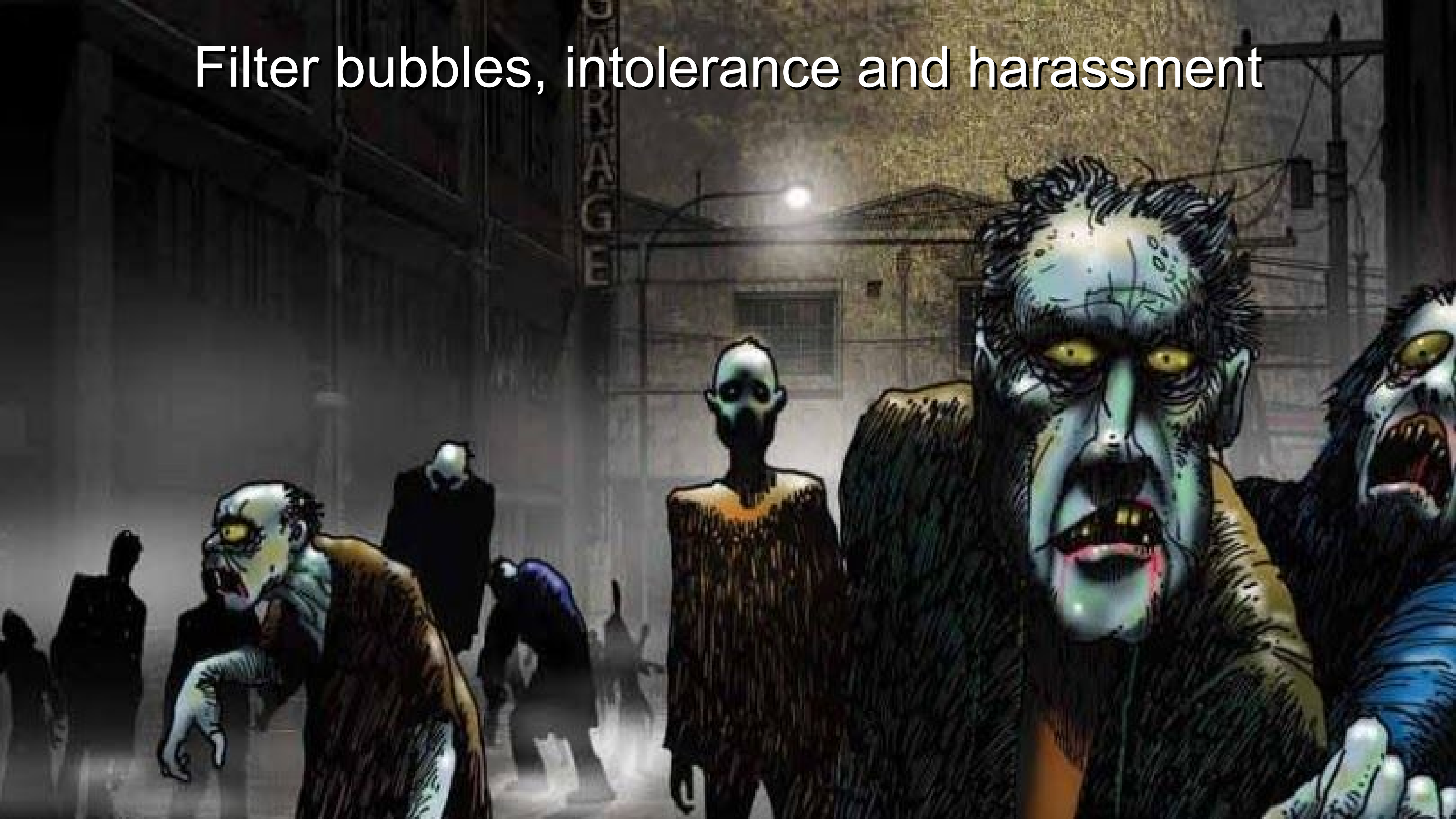
jigsaw.google.com

A Utopian promise of the Internet?

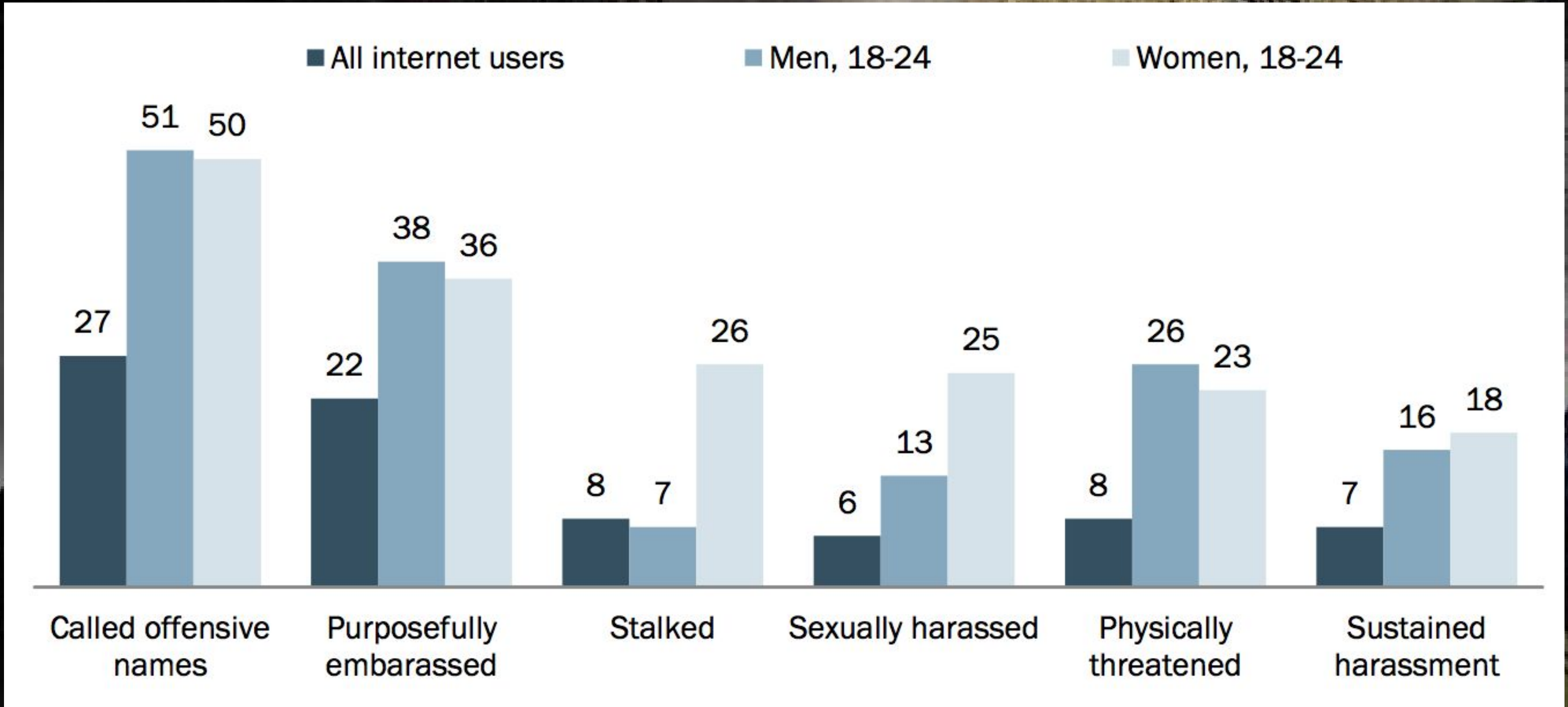


Conversation on the Internet?

Filter bubbles, intolerance and harassment



Harassment comes in many forms



Source: M. Duggan. *Online harassment.* Pew Research Center, 2014.

Pew 2017 Harassment Report
(US-based Internet users):

After witnessing the harassment of others:

- 27% refrained from posting online
- 13% stopped using an online service

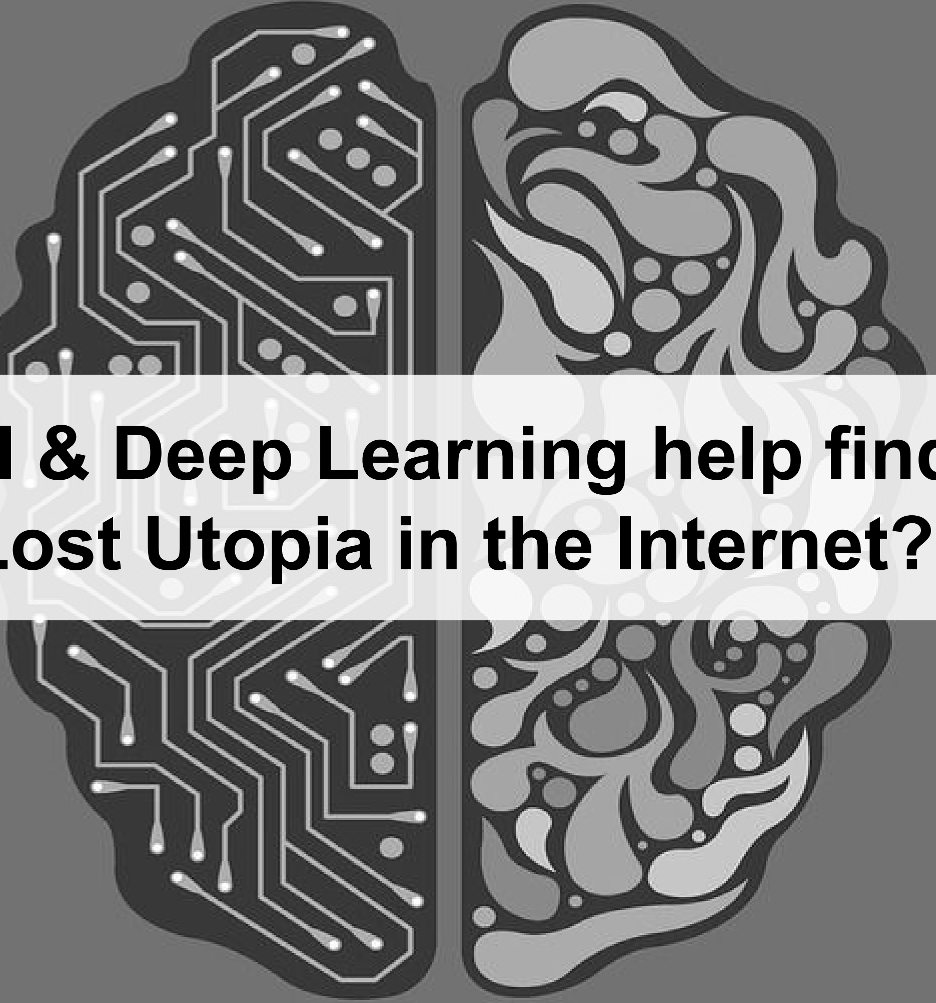
41% personally subjected to online harassment

http://www.pewinternet.org/2017/07/11/online-harassment-2017/

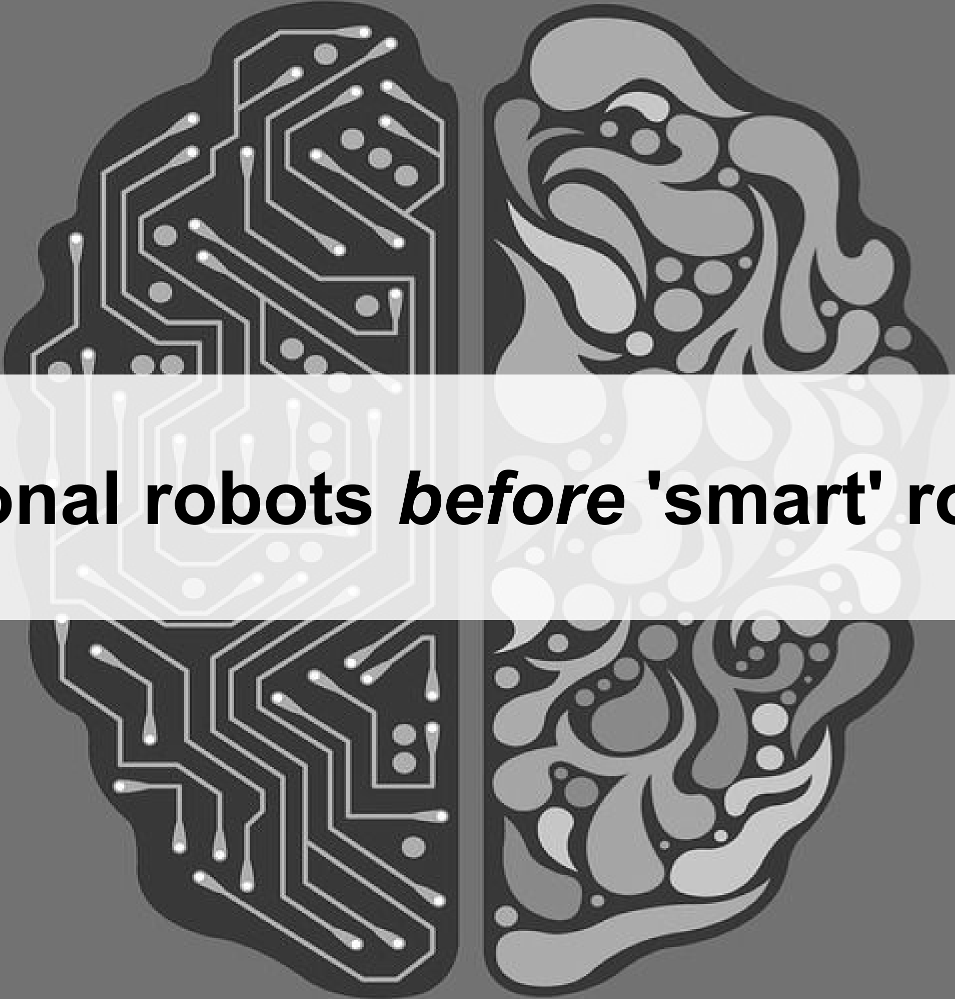
Despite user engagement as a key to success, many news platforms and blogs turn off comments.

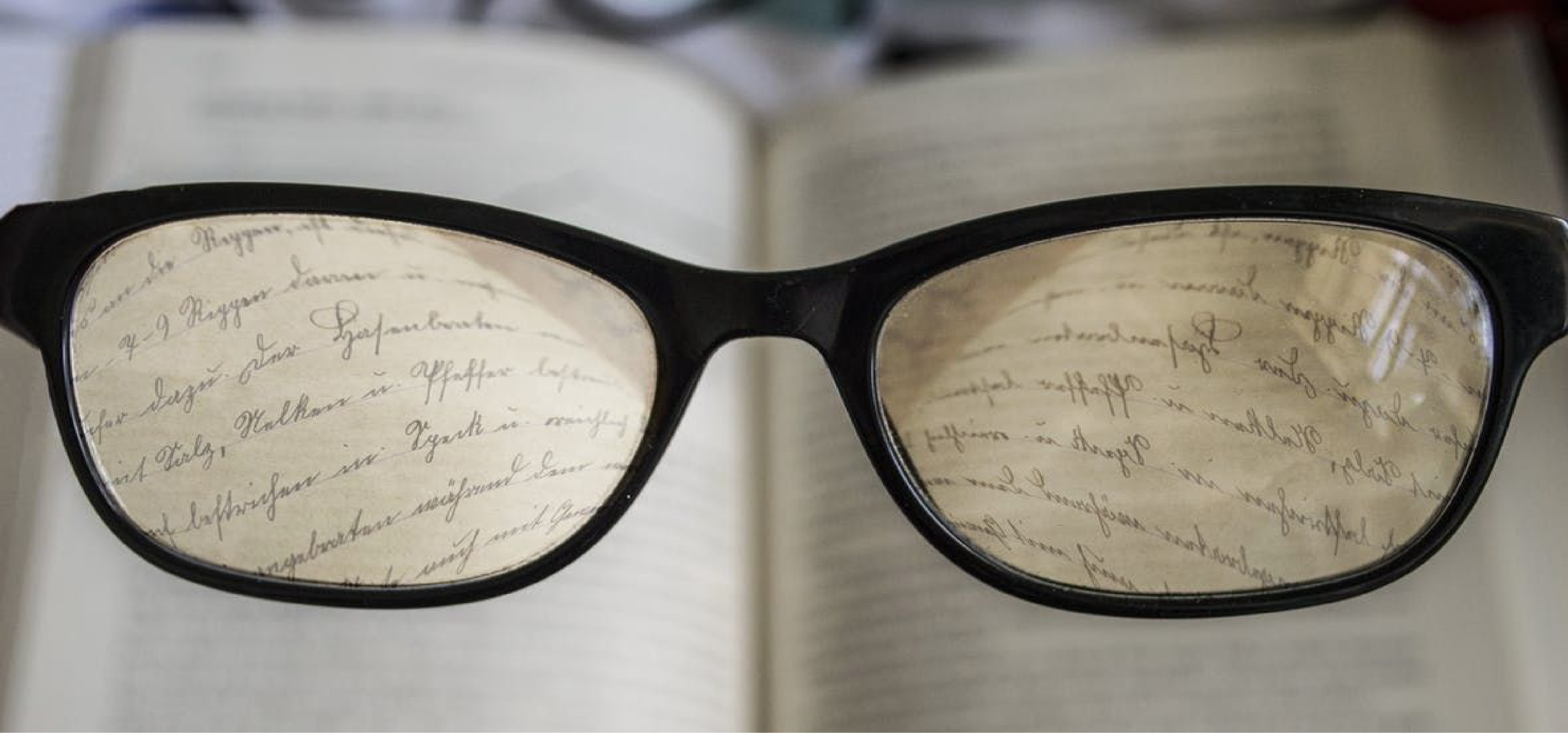
The eternal desire for good conversation

Can AI & Deep Learning help find that Lost Utopia in the Internet?



Emotional robots *before* 'smart' robots?





ML to support conversations?

What should the ML look for?
What do you want to find?

A starting point:

comments that are likely to make people leave the discussion

(lets call this *Toxicity*)

NOT SURE IF FUNNY
OR JUST OFFENSIVE

What a fat pig!

What a fat pig!

Found It!!!
Congratulations, it only took you 65298 seconds
www.jolyon.co.uk

Unintended Bias

False "toxic" positives

A naively trained model on will have some strong unintended biases illustrated by these false-positive examples...

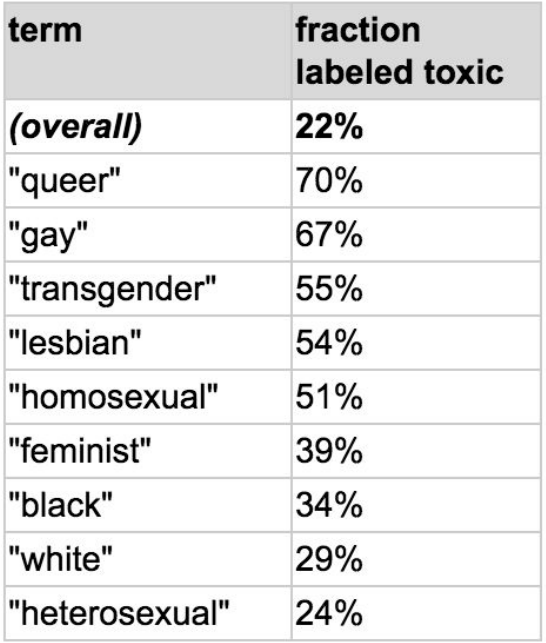
Comment	Toxicity score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam.	0.46

How did this happen?

ML over-generalizes due to:

- Insufficient data
- The 'real' distribution is skewed

The model is not able to distinguish toxic from non-toxic uses of many identity words (and some others too, e.g. donkey)



term	fraction labeled toxic
(overall)	**22%**
"queer"	70%
"gay"	67%
"transgender"	55%
"lesbian"	54%
"homosexual"	51%
"feminist"	39%
"black"	34%
"white"	29%
"heterosexual"	24%

Unintended Model Bias vs Unfairness

- **Model: Unintended Bias** (A subset of examples has an unintended score distirbution)
  Application: Unfairness (Unfair impact on people)

- Unintended bias can easily lead to unfair applications.

- **Every application of ML needs to consider the potential impact of unintended bias on the application's impact on society (fairness, inclusivity, etc).**
  - Unintended bias can lead to behaviour that increases, or decreases, the prevalence of mentions of an identity group (or it may have not effect); e.g. human pre-moderation, post-moderation, and batch moderation respectively.

How to measure Unintended Bias?

How good is the model at distinguishing good from bad examples? (ROC-AUC)
AUC (for a given test set) = Given two examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.

*Pinned AUC (for a given term, **t**, in a test set) =*
*AUC(all N examples with **t** & N representative examples from the test set)*

Pinned AUC < AUC if the model gives unusually high (or low) scores to examples containing the term **t**. PinnedAUCΔ = if AUC > PinnedAUC then (AUC - PinnedAUC) else 0.

Unintended bias for identity terms = ∑ PinnedAUCΔ(**t**, **s**), for each identity term **t** in a balanced test set **s** (e.g. a synthetic test set based on templates with identity terms)

Mitigating unintended bias: re-balance the dataset

Where to get non-toxic examples about terms that are most frequently in toxic comments?

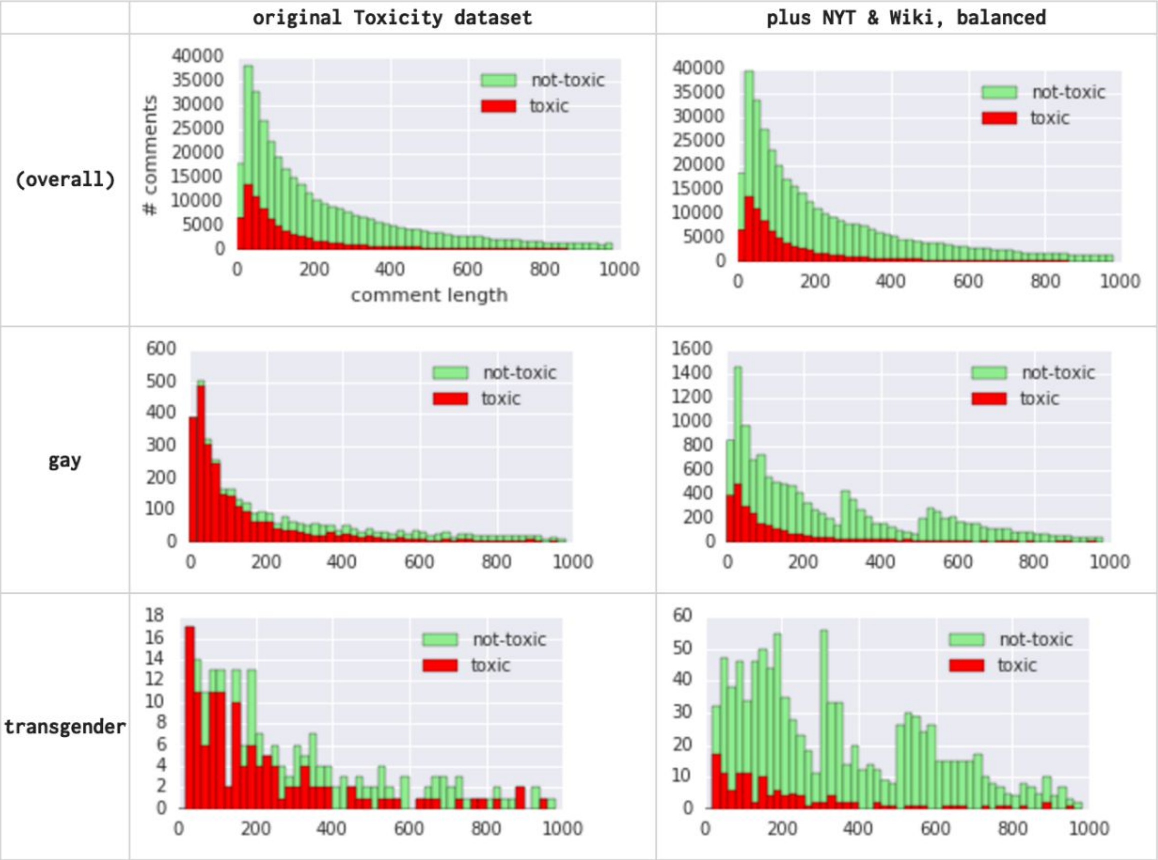
- Wikipedia Article Pages! (or other reviewed sources; reviewed comments, articles, etc)
- Re-balance the examples for each term (by length, this is important)

Potential issues:

- Text in article pages is not the same as text in comments, will this work?
- Will you have enough examples?

Mitigating unintended bias? re-balance the data

term	fraction labeled toxic
(overall)	**22%**
"queer"	70%
"gay"	67%
"transgender"	55%
"lesbian"	54%
"homosexual"	51%
"feminist"	39%
"black"	34%
"white"	29%
"heterosexual"	24%



False positives - some improvement

Comment	Old	New
The Gay and Lesbian Film Festival starts today.	0.82	0.01
Being transgender is independent of sexual orientation.	0.52	0.05
A Muslim is someone who follows or practices Islam.	0.46	0.13

Overall AUC for old and new classifiers within noise of retraining.

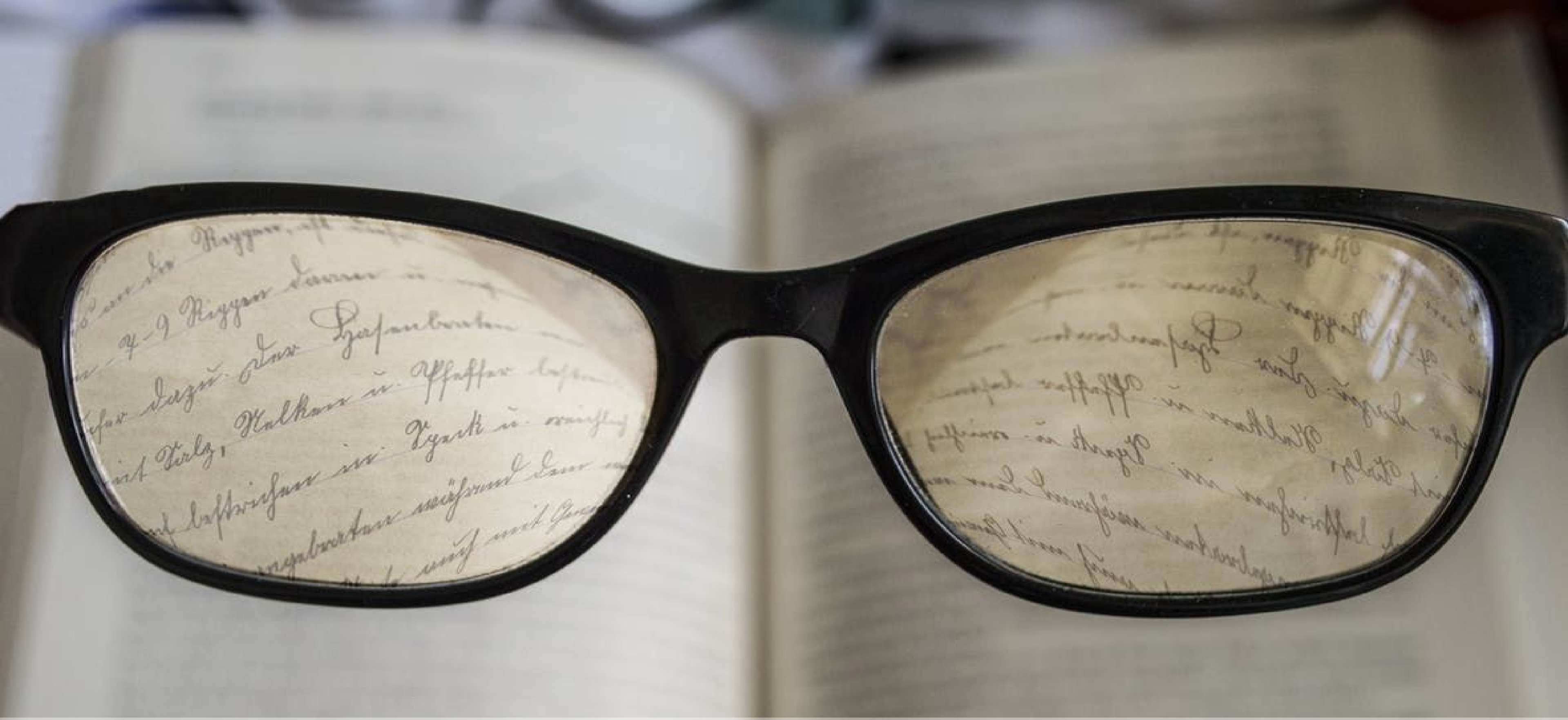
Many open questions

- Where to get a balances test set of identity terms?
- Should we be doing a squared error calculation?

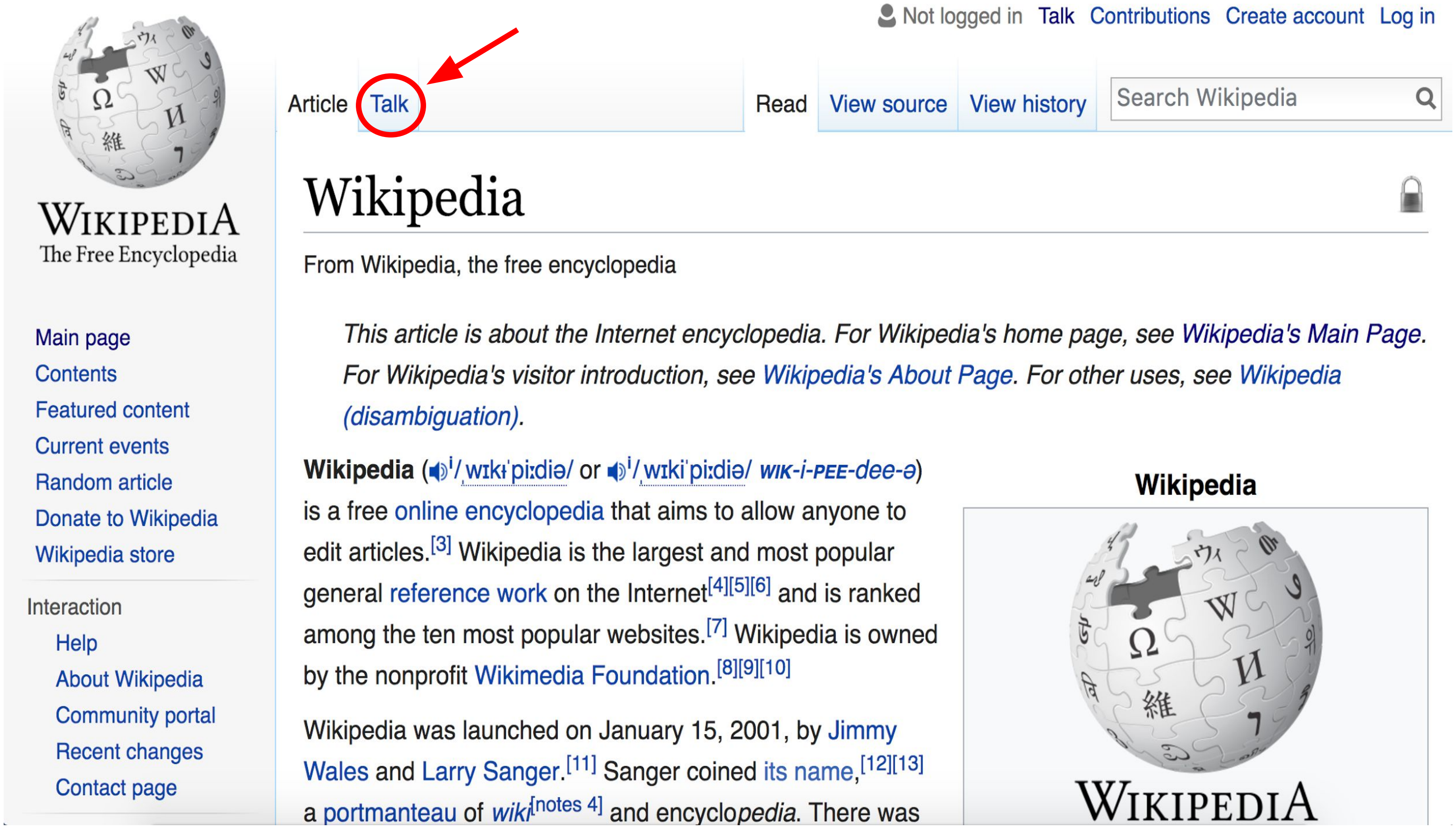
Adversarial examples from public demos help a lot too.

But this does not make a 'perfect' model - that does not exist, a lot more hard work is needed here, and this will be a challenge for a long time.

https://github.com/conversationai/unintended-ml-bias-analysis
(built on Wikipedia, includes ML models, and mitigation methods)



Some examples of ML to support conversation



Wikipedia

From Wikipedia, the free encyclopedia

This article is about the Internet encyclopedia. For Wikipedia's home page, see Wikipedia's Main Page. For Wikipedia's visitor introduction, see Wikipedia's About Page. For other uses, see Wikipedia (disambiguation).

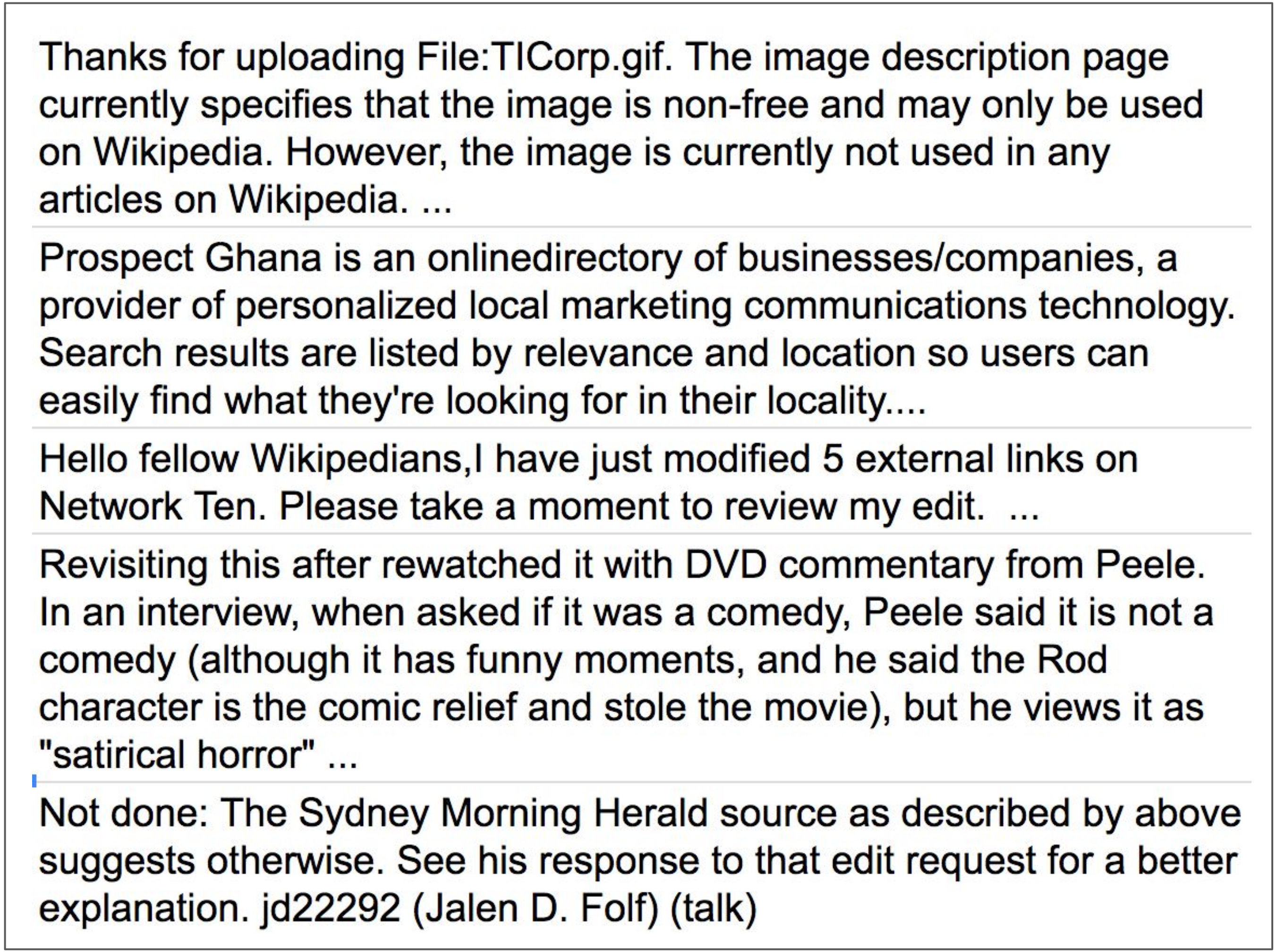
Wikipedia

Wikipedia (🔊i/ˌwɪkɪˈpiːdiə/ or 🔊i/ˌwɪkiˈpiːdiə/ *WIK-i-PEE-dee-ə*) is a free online encyclopedia that aims to allow anyone to edit articles.[3] Wikipedia is the largest and most popular general reference work on the Internet[4][5][6] and is ranked among the ten most popular websites.[7] Wikipedia is owned by the nonprofit Wikimedia Foundation.[8][9][10]

Wikipedia was launched on January 15, 2001, by Jimmy Wales and Larry Sanger.[11] Sanger coined its name,[12][13] a portmanteau of *wiki*[notes 4] and encyclo*pedia*. There was

Chronologically

First few comments of 11,365 from Wikipedia on September 4th, 2017



Thanks for uploading File:TICorp.gif. The image description page currently specifies that the image is non-free and may only be used on Wikipedia. However, the image is currently not used in any articles on Wikipedia. ...

Prospect Ghana is an onlinedirectory of businesses/companies, a provider of personalized local marketing communications technology. Search results are listed by relevance and location so users can easily find what they're looking for in their locality....

Hello fellow Wikipedians,I have just modified 5 external links on Network Ten. Please take a moment to review my edit. ...

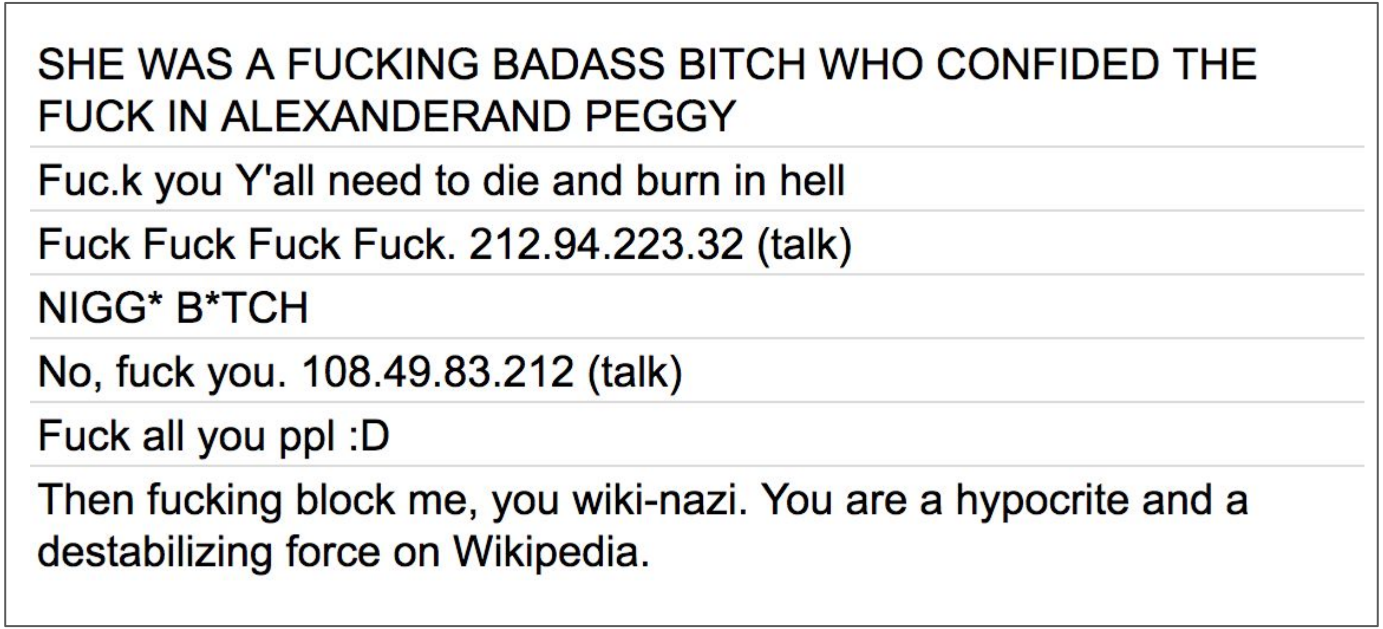
Revisiting this after rewatched it with DVD commentary from Peele. In an interview, when asked if it was a comedy, Peele said it is not a comedy (although it has funny moments, and he said the Rod character is the comic relief and stole the movie), but he views it as "satirical horror" ...

Not done: The Sydney Morning Herald source as described by above suggests otherwise. See his response to that edit request for a better explanation. jd22292 (Jalen D. Folf) (talk)

medium.com/the-false-positive/better-discussions-with-imperfect-models-91558235d442

By toxicity (probability)

Top few of 11,365 from Wikipedia on September 4th, 2017



SHE WAS A FUCKING BADASS BITCH WHO CONFIDED THE FUCK IN ALEXANDERAND PEGGY

Fuc.k you Y'all need to die and burn in hell

Fuck Fuck Fuck Fuck. 212.94.223.32 (talk)

NIGG* B*TCH

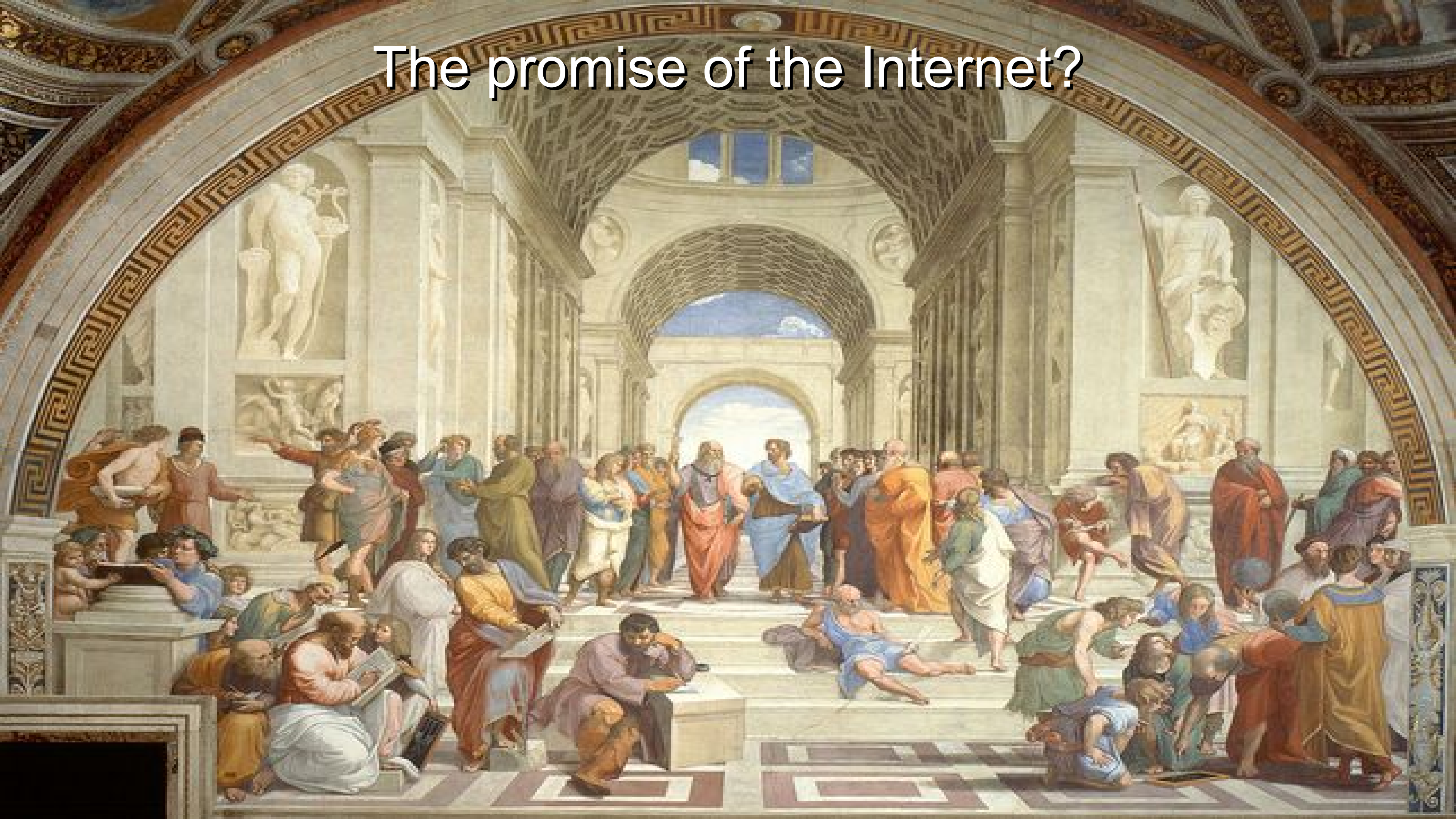
No, fuck you. 108.49.83.212 (talk)

Fuck all you ppl :D

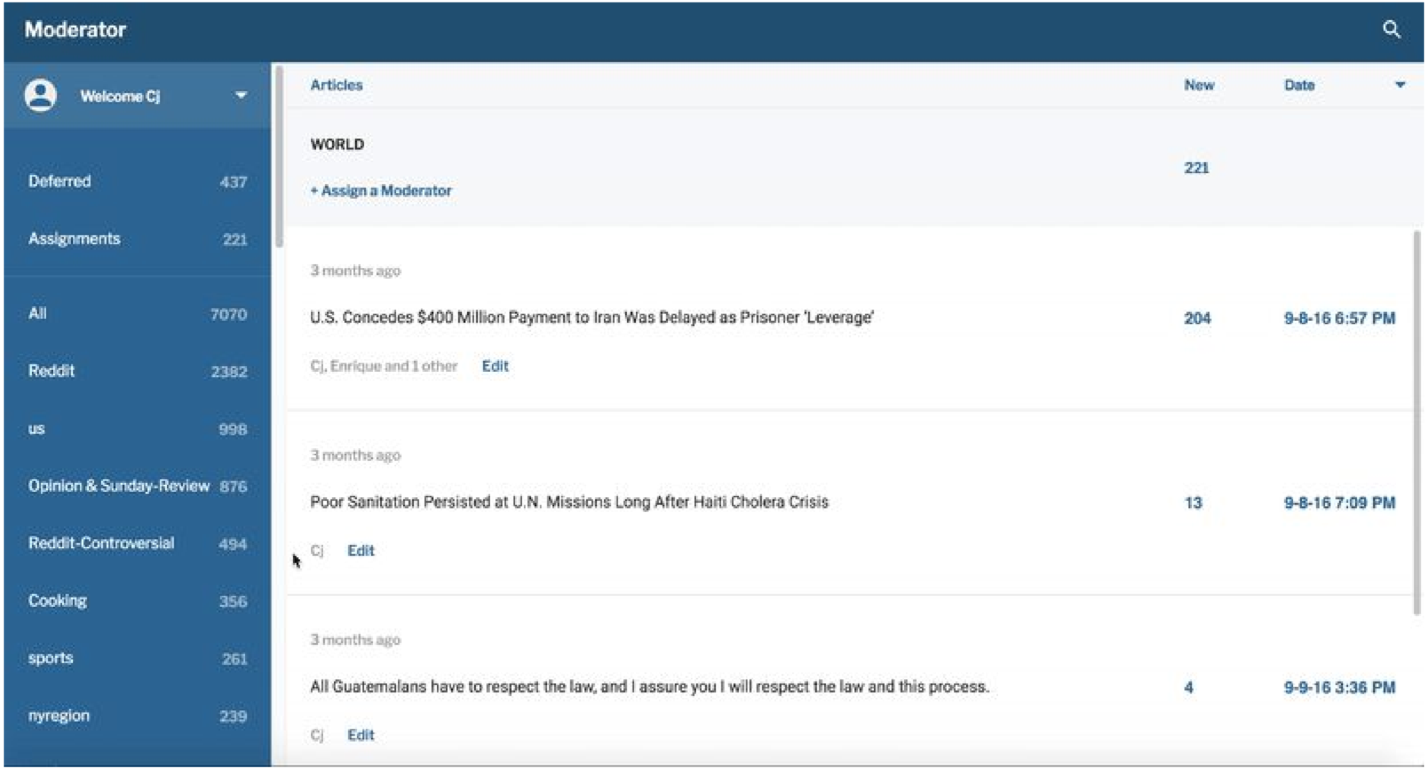
Then fucking block me, you wiki-nazi. You are a hypocrite and a destabilizing force on Wikipedia.

medium.com/the-false-positive/better-discussions-with-imperfect-models-91558235d442

The promise of the Internet?



What might ML do to help? Assisted Moderation



Moderator

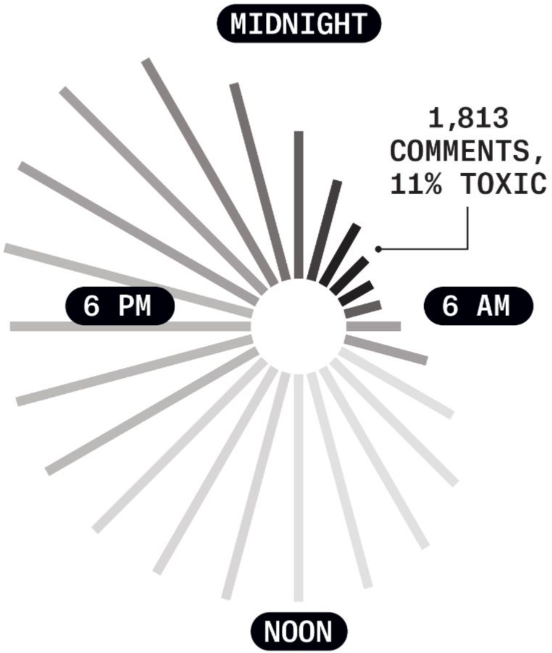
Welcome Cj

Deferred	437
Assignments	221
All	7070
Reddit	2382
us	998
Opinion & Sunday-Review	876
Reddit-Controversial	494
Cooking	356
sports	261
nyregion	239

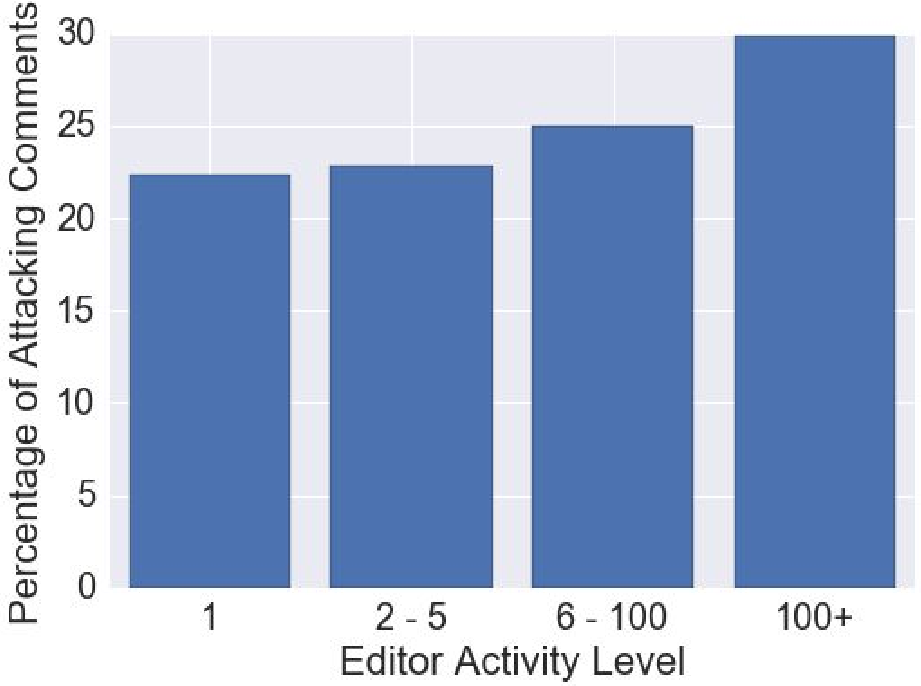
Articles	New	Date
WORLD + Assign a Moderator	221	
3 months ago U.S. Concedes $400 Million Payment to Iran Was Delayed as Prisoner 'Leverage' Cj, Enrique and 1 other Edit	204	9-8-16 6:57 PM
3 months ago Poor Sanitation Persisted at U.N. Missions Long After Haiti Cholera Crisis Cj Edit	13	9-8-16 7:09 PM
3 months ago All Guatemalans have to respect the law, and I assure you I will respect the law and this process. Cj Edit	4	9-9-16 3:36 PM

Open Source Code @ github.com/conversationai/conversationai-moderator

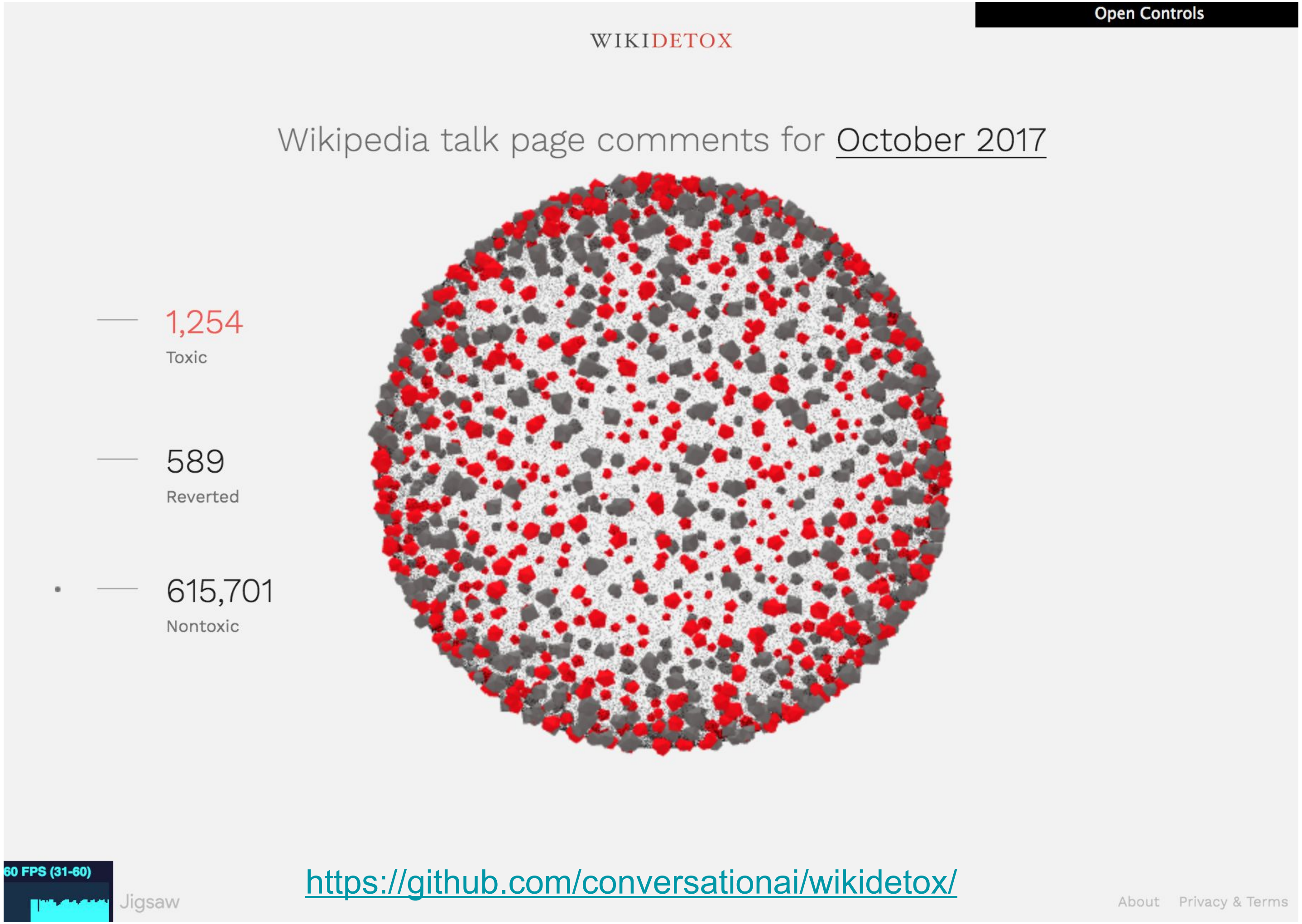
What might ML do to help? Understanding



https://www.wired.com/2017/08/internet-troll-map/



Open Source Code @ github.com/ewulczyn/wiki-detox



Wikipedia talk page comments for October 2017

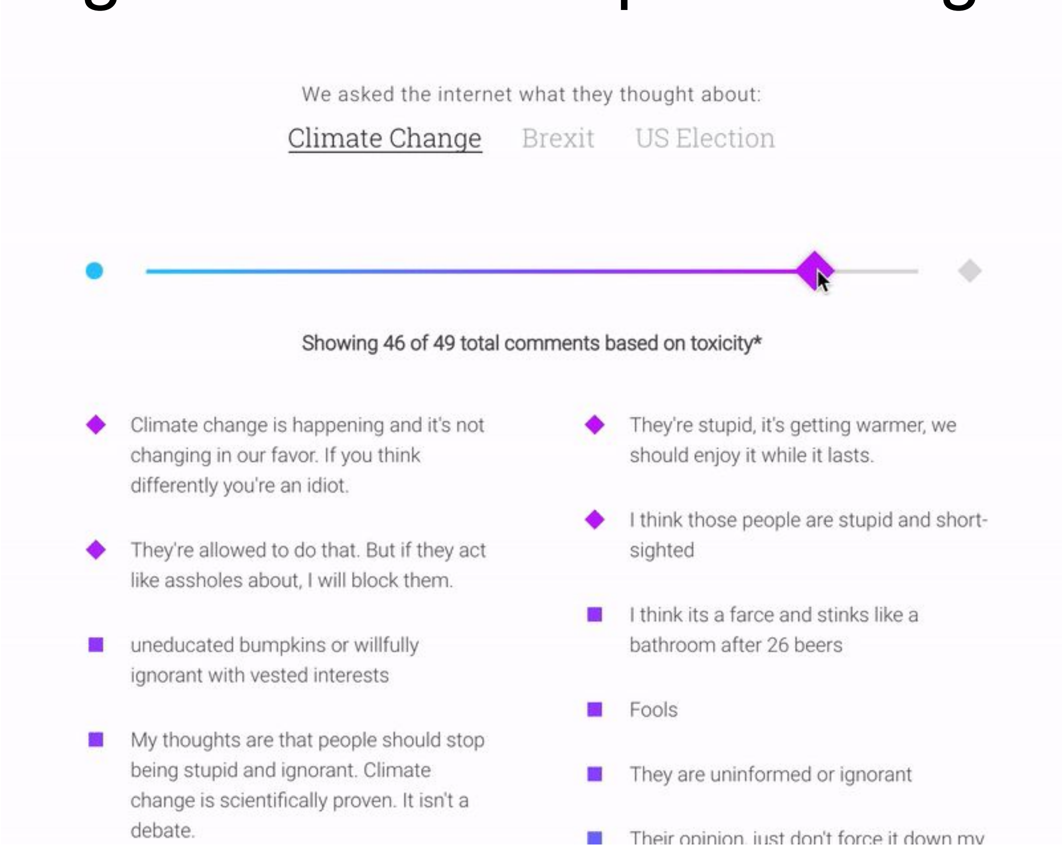
1,254
Toxic

589
Reverted

615,701
Nontoxic

https://github.com/conversationai/wikidetox/

What might ML do to help? Viewing comments



We asked the internet what they thought about:

Showing 46 of 49 total comments based on toxicity*

- ◆ Climate change is happening and it's not changing in our favor. If you think differently you're an idiot.
- ◆ They're allowed to do that. But if they act like assholes about, I will block them.
- ■ uneducated bumpkins or willfully ignorant with vested interests
- ■ My thoughts are that people should stop being stupid and ignorant. Climate change is scientifically proven. It isn't a debate.
- ◆ They're stupid, it's getting warmer, we should enjoy it while it lasts.
- ◆ I think those people are stupid and short-sighted
- ■ I think its a farce and stinks like a bathroom after 26 beers
- ■ Fools
- ■ They are uninformed or ignorant
- ■ Their opinion, just don't force it down my

perspectiveapi.com

What might ML do to help? Authorship

Demo at: perspectiveapi.com
Open Source Code @ github.com/conversationai/perspectiveapi-authorship-demo
It will get lots of things wrong, but when you trick it, it helps correct biases, and it helps me, so please do!

What a fat pig!

Conversation Context

The models we have developed so far largely ignore the context of a comment (they do use the context of words within a comment).

But how important is that? How can we measure it?

First step interpret conversations on Wikipedia

- Conversations happen on Talk Pages.
- Talk pages are a series of revisions... talk pages need interpretation.

A tool for further research: structured conversations

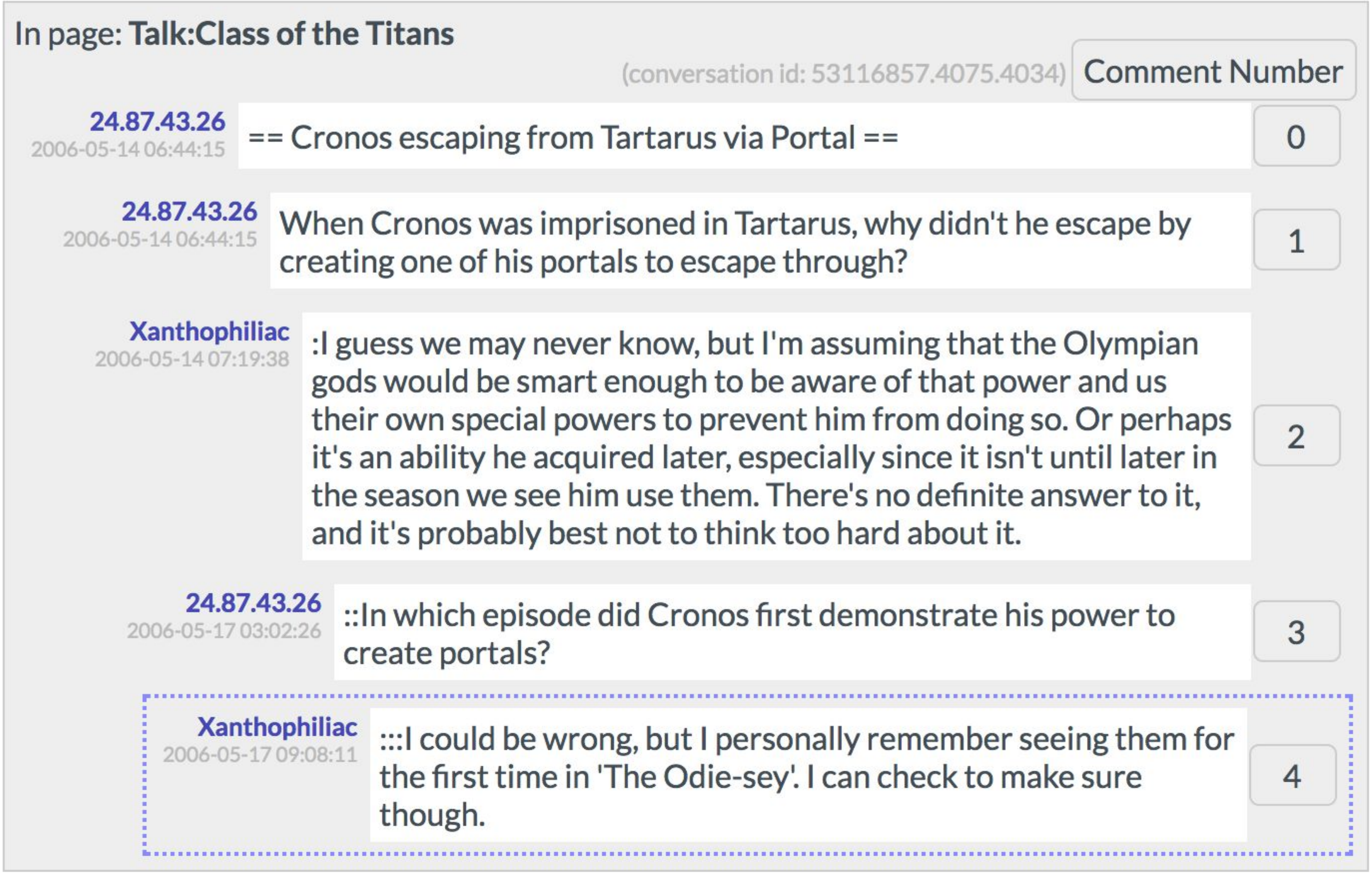
Conversation Reconsturction on Wikipedia

Snapshots: parse a snapshot of the talk pages, and construct the conversation from that snapshot.

- Loses authorship, history of actions, modifications to comments, and removed comments (common in abuse cases)

Diff History: Creation of a section, Additions, Modifications, Removals.

- Parsing is harder, need to process pages edits sequentially
- Can reply conversation history over time
- Higher fidelity & keeps authorship, modifications, and deletions.



In page: **Talk:Class of the Titans**

(conversation id: 53116857.4075.4034)

24.87.43.26
2006-05-14 06:44:15
== Cronos escaping from Tartarus via Portal ==
0

24.87.43.26
2006-05-14 06:44:15
When Cronos was imprisoned in Tartarus, why didn't he escape by creating one of his portals to escape through?
1

Xanthophiliac
2006-05-14 07:19:38
:I guess we may never know, but I'm assuming that the Olympian gods would be smart enough to be aware of that power and us their own special powers to prevent him from doing so. Or perhaps it's an ability he acquired later, especially since it isn't until later in the season we see him use them. There's no definite answer to it, and it's probably best not to think too hard about it.
2

24.87.43.26
2006-05-17 03:02:26
::In which episode did Cronos first demonstrate his power to create portals?
3

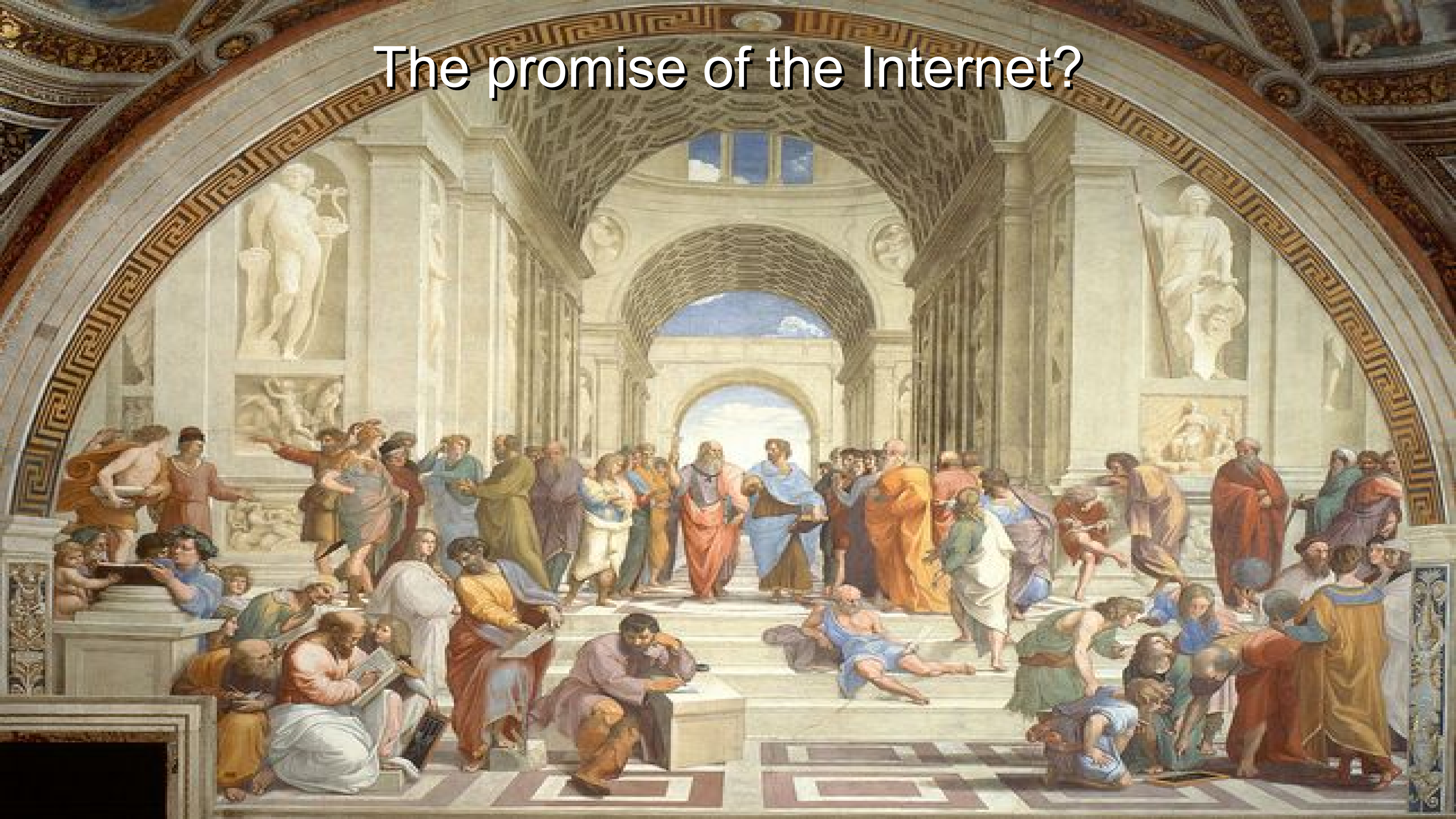
Xanthophiliac
2006-05-17 09:08:11
:::I could be wrong, but I personally remember seeing them for the first time in 'The Odie-sey'. I can check to make sure though.
4

Opens New Questions:

- How important is the conversational context to human judgements of toxicity?
  - In theory it's critical, but in practice it seems less so, but now we can quantify it!
- Are there early indicators for conversations becoming toxic?
- Are there people who seem to know how to not have toxic conversations?
  - What can we learn from their conversations?
- Can we design and test ways to help conversations on Wikipedia?

This is ongoing joint work with Wikimedia Research & Cornell Techn
Cristian Danescu, Yiqing Hua, Dario Taraborelli Nithum Thain

The promise of the Internet?



In summary

- **Toxic conversation online seems like a wicked problem**
  - Online harassment is pervasive, silencing and siloing

- **ML provides new ways to think about online conversation**
  - And new challenges: e.g. unintended bias & fair application of ML
  - ML enables new UX to assist people: for curators of conversations, authors, and viewers.
  - Opens new questions for the study of online conversation

Our public code and data is at: conversationai.github.io

If you don't want to build your own models: perspectiveapi.com

Much of this work is part of the WikiDetox project as well as ongoing research on the nature and impact of harassment in Wikipedia discussion spaces – part of a collaboration between Jigsaw, Cornell University, and the Wikimedia Foundation