# Open, collaborative, reproducible research to support free knowledge

*I Congresso Científico Brasileiro da Wikipédia* • October 14, 2016

Dario Taraborelli
@readermeter


# Outline

Research at the Wikimedia Foundation

Wikimedia Research: current priorities

A platform for open, collaborative, reproducible research

# Outline

Research at the Wikimedia Foundation

Wikimedia Research: current priorities

A platform for open, collaborative, reproducible research

# Outline

Research at the Wikimedia Foundation

Wikimedia Research: current priorities

A platform for open, collaborative, reproducible research

# Research at the Wikimedia Foundation

# A brief history of research at WMF

Early years

2010 Research committee (RCom), the Research Index

2011 First research roles

2011 Wikimedia Summer of Research

2012 Research microteams at the Foundation

2013 Research and Analytics

2015 Research department

# A brief history of research at WMF

| | |
|---|---|
| | Early years |
| 2010 | Research committee (RCom), the Research Index |
| 2011 | First research roles |
| 2011 | Wikimedia Summer of Research |
| 2012 | Research microteams at the Foundation |
| 2013 | Research and Analytics |
| 2015 | Research department |

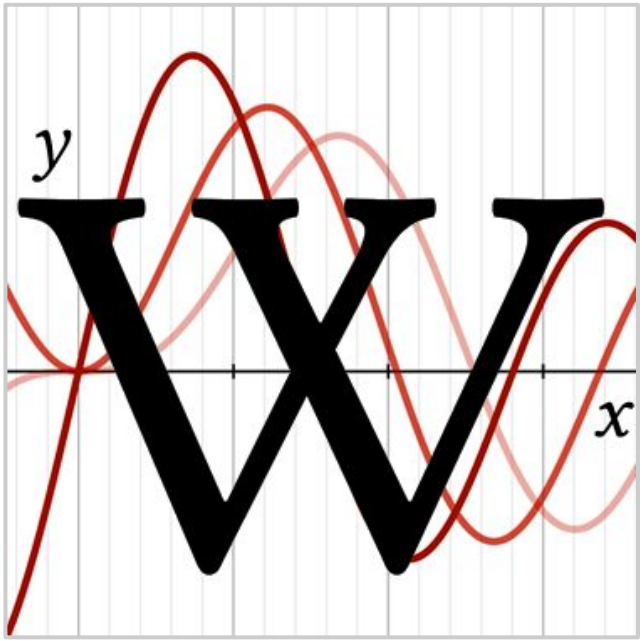
# A brief history of research at WMF

| | |
|---|---|
| | Early years |
| 2010 | Research committee (RCom), the Research Index |
| 2011 | First research roles |
| 2011 | Wikimedia Summer of Research |
| 2012 | Research microteams at the Foundation |
| 2013 | Research and Analytics |
| 2015 | Research department |

# A brief history of research at WMF

| | |
|---|---|
| | Early years |
| 2010 | Research committee (RCom), the Research Index |
| 2011 | First research roles |
| 2011 | Wikimedia Summer of Research |
| 2012 | Research microteams at the Foundation |
| 2013 | Research and Analytics |
| 2015 | Research department |

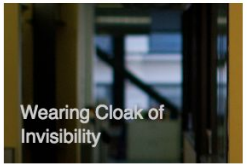
# Wikimedia Research




https://www.mediawiki.org/wiki/Wikimedia_Research

## Research

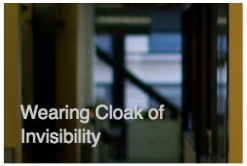
**Dario Taraborelli**
Director, Head of Research




**Nathaniel Schaaf**
Software Engineer
(International)

## Research and Data

**Aaron Halfaker**
Principal Research Scientist




**Morten Warncke-Wang**
Research Fellow

**Robert West**
Research Fellow

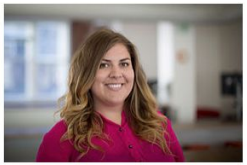
**Ellery Wulczyn**
Data Scientist

**Erik Zachte**
Data Analyst (International
contractor)

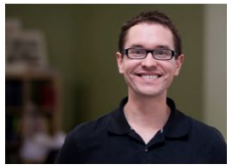
**Leila Zia**
Senior Research Scientist

## Design Research

**Abigail Ripstra**
Lead Design Research
Manager



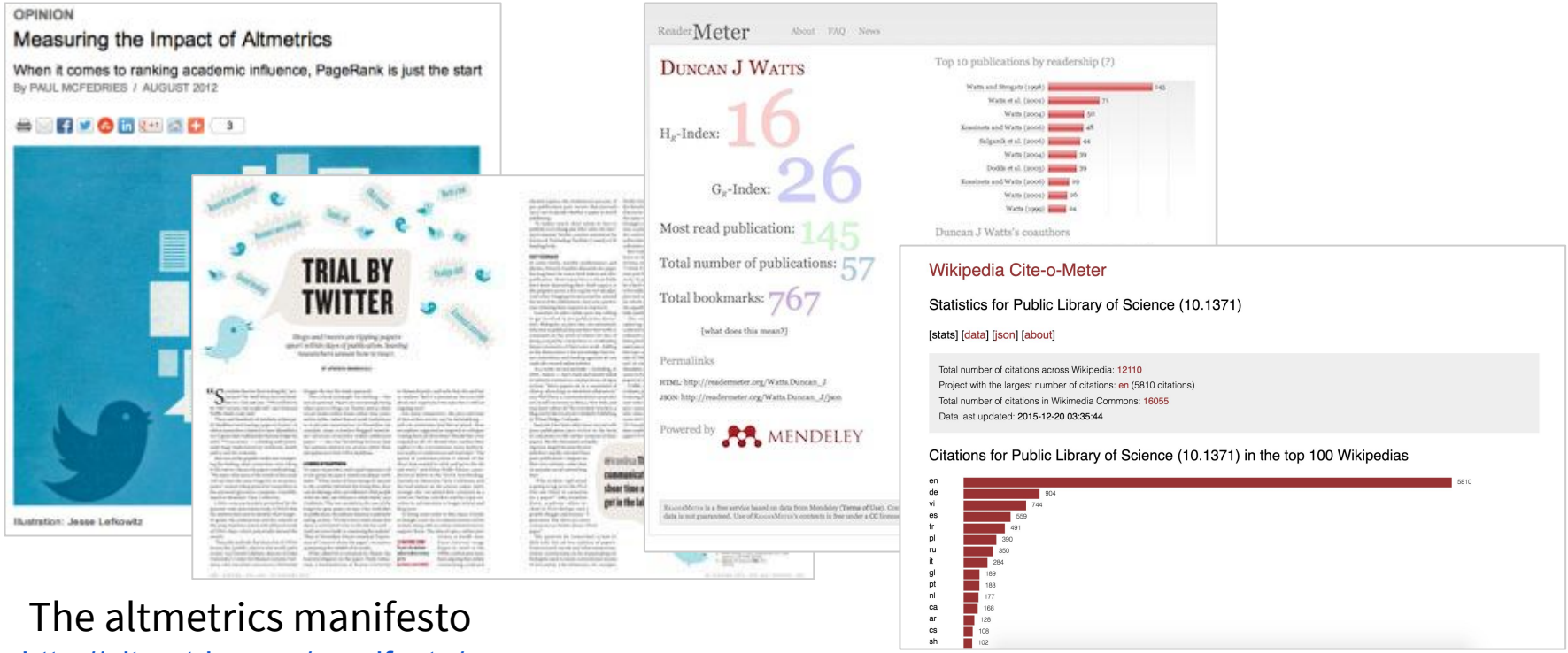
**Samantha Becker**
Participant Recruiter
(Contractor)



**Jonathan Morgan**
Senior Design Researcher

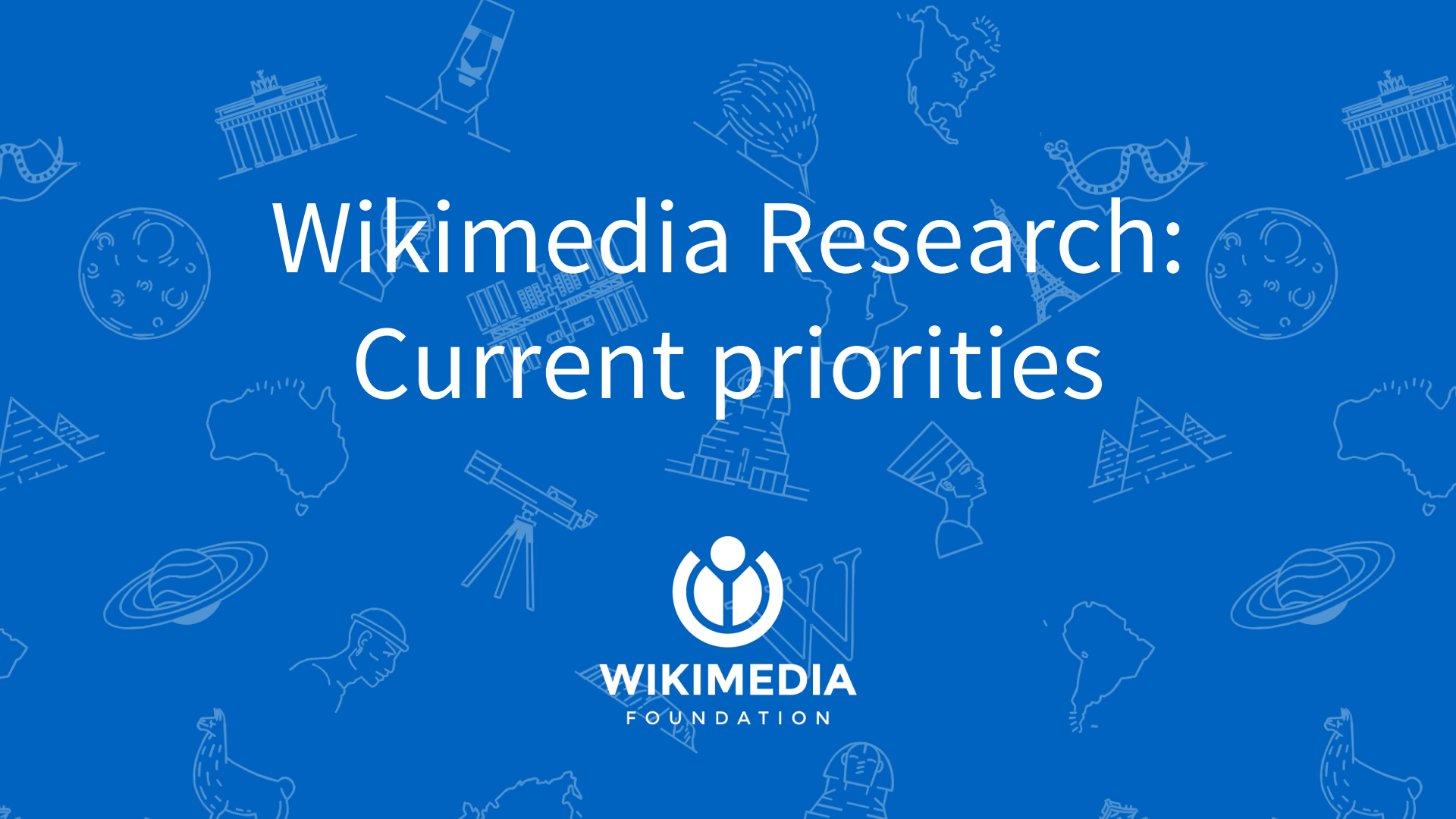
https://www.mediawiki.org/wiki/Wikimedia_Research

# altmetrics




The altmetrics manifesto
http://altmetrics.org/manifesto/

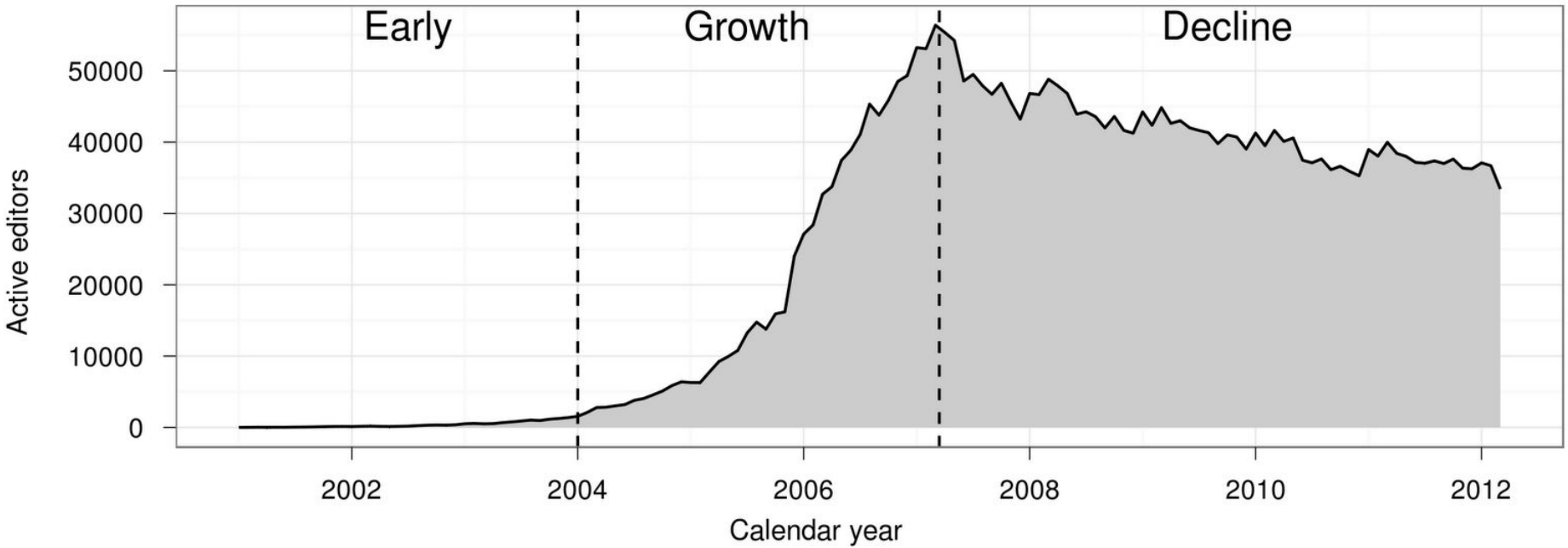
# Wikimedia Research: Current priorities



# Artificial intelligence as a service

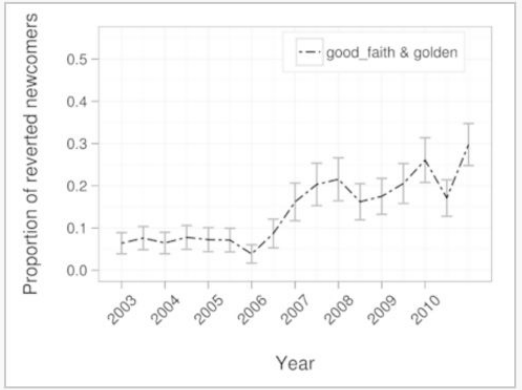
WIKIMEDIA
FOUNDATION

1. *Socially aware* quality control

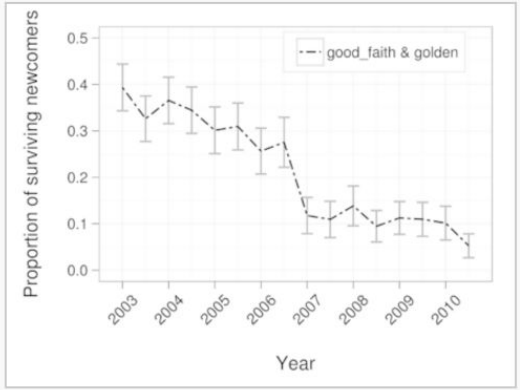



https://meta.wikimedia.org/wiki/Research:The_Rise_and_Decline

# first-time editors are *as good as they were* in 2004-05, but their *retention* sharply declined




**Rejection of desirable newcomers.** The proportion of desirable newcomers who are reverted in their first session of editing is plotted over time.




**Survival of desirable newcomers.** The proportion of desirable newcomers who edit for at least two months is plotted over time.

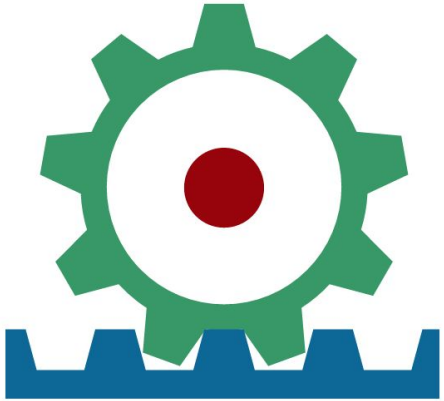
Wikipedia is a firehose
Bad edits must be reverted
Cost of quality control work must be minimized

*What about new contributors?*
*How to build socially-aware quality control systems?*

https://www.mediawiki.org/wiki/File:The_people's_classifier_--_Research_Showcase_(May,_2015).pdf

# Correcting the side effects of quality control

The Objective Revision Evaluation Service (ORES) is a web service that provides machine learning as a service for Wikimedia Projects. The system is designed to help automate critical wiki-work -- for example, vandalism detection and removal. This service is developed as part of the Revision scoring as a service research project.



## Scores API

ORES is intended to be used as a source of information by tool developers. To access ORES, scores, a simple RESTful API is provided. There are two versions of the scoring API that differ slightly in their behavior. Version 2 provides access to model info in a scoring request. Version 1 is preserved for backwards compatibility.

http://ores.wikimedia.org
https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/

# The ORES API

**Revision as of 21:52, 12 January 2015 (edit) (undo)**
Blank123456789 (talk | contribs)
Next edit →

**Line 11:**

==Evaluation of intelligence==

\+ "Intelligence" is hard to define, whether in dogs, other animals, or humans. The ability to learn quickly might be taken as a sign of [[intelligence (trait)|intelligence]], but such evidence must be interpreted with care, because learning speed may be affected by such things as the effectiveness of the rewards used in training or the motivation or activity level of the dog. For example, some breeds, such as [[Siberian Husky|Siberian Huskies]], are said to be not particularly rewarded by pleasing their owners, but quickly learn to escape from yards or catch small animals, often using ingenious ways of doing both. **LLAMAS GROW ON TREES**


```
{
  "642215410": {
    "prediction": true,
    "probability": {
      "false": 0.07561753062984156,
      "true": 0.9243824693701584
    }
  }
}
```

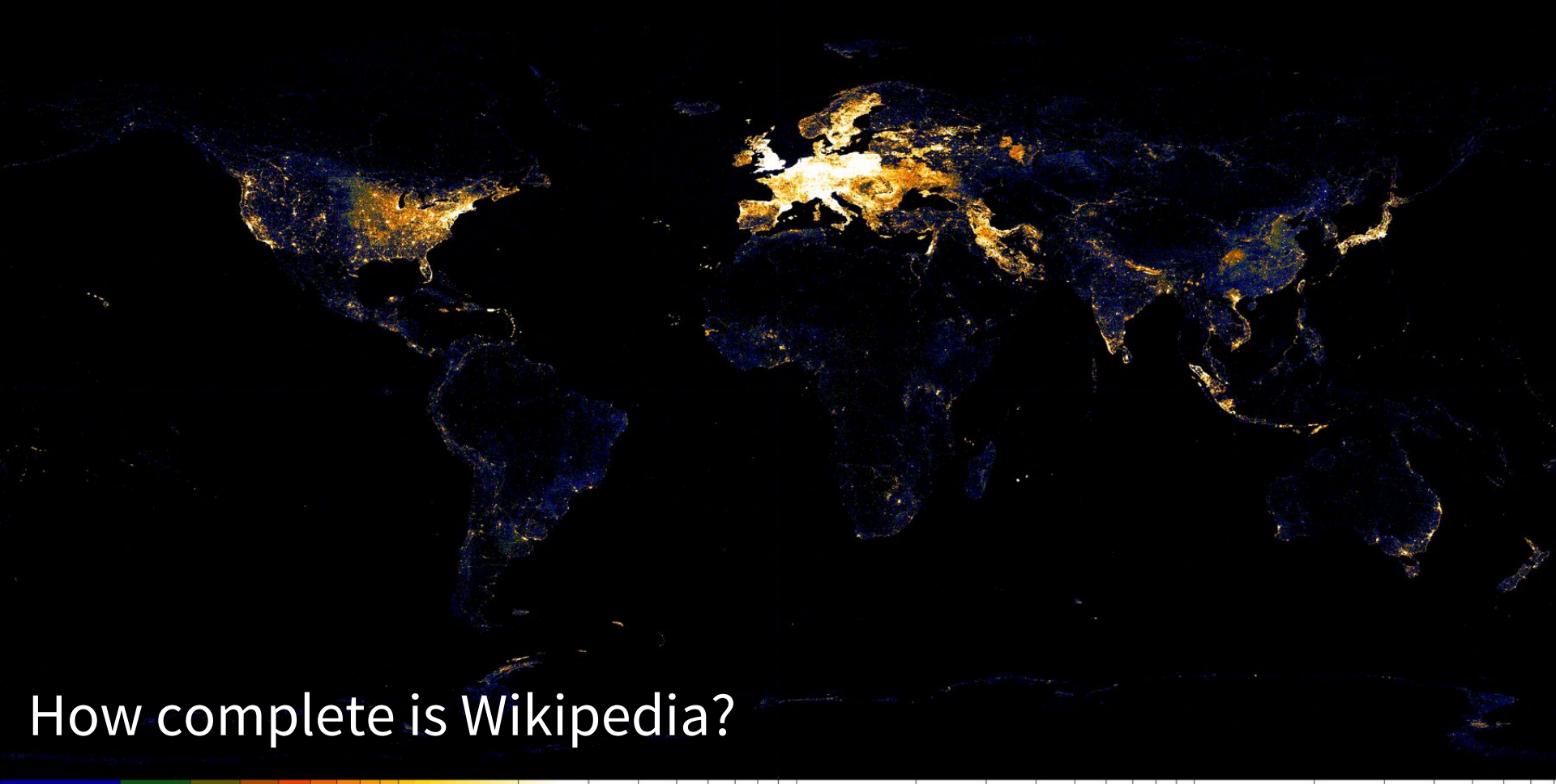

https://ores.wikimedia.org/scores/enwiki/damaging/642215410
https://en.wikipedia.org/w/index.php?diff=prev&oldid=642215410

# The ORES API


http://ores.wikimedia.org
https://meta.wikimedia.org/wiki/ORES
Image credits: Mun May Tee-Galloway

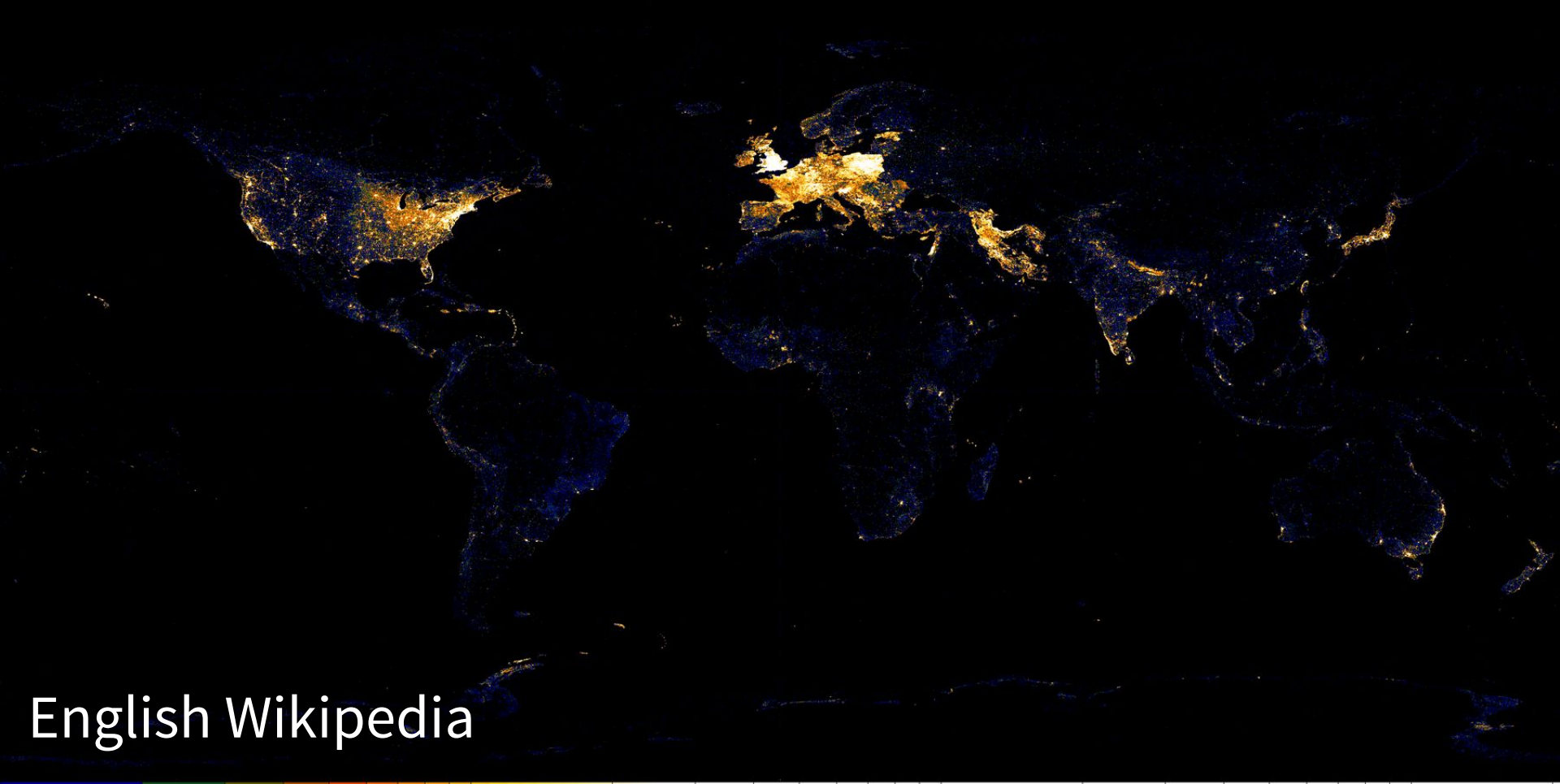
| context | edit quality | | | article quality |
|---|---|---|---|---|
| | damaging | goodfaith | reverted | wp10 |
| arwiki Arabic Wikipedia | | | ✓ | |
| cswiki Czech Wikipedia | | | ✓ | |
| dewiki German Wikipedia | | | ✓ | |
| enwiki English Wikipedia | ✓ | ✓ | ✓ | ✓ |
| enwiktionary English Wiktionary | | | ✓ | |
| eswiki Spanish Wikipedia | | | ✓ | |
| etwiki Estonian Wikipedia | | | ✓ | |
| fawiki Persian Wikipedia | ✓ | ✓ | ✓ | |
| frwiki French Wikipedia | | | ✓ | ✓ |
| hewiki Hebrew Wikipedia | | | ✓ | |
| huwiki Hungarian Wikipedia | | | ✓ | |
| idwiki Indonesian Wikipedia | | | ✓ | |
| itwiki Italian Wikipedia | | | ✓ | |
| nlwiki Dutch Wikipedia | ✓ | ✓ | ✓ | |
| nowiki Norwegian Wikipedia | | | ✓ | |
| plwiki Polish Wikipedia | ✓ | ✓ | ✓ | |
| ptwiki Portuguese Wikipedia | ✓ | ✓ | ✓ | |
| ruwiki Russian Wikipedia | ✓ | ✓ | ✓ | ✓ |
| svwiki Swedish Wikipedia | | | ✓ | |
| trwiki Turkish Wikipedia | ✓ | ✓ | ✓ | |
| ukwiki Ukranian Wikipedia | | | ✓ | |
| viwiki Vietnamese Wikipedia | | | ✓ | |
| wikidatawiki Wikidata | ✓ | ✓ | ✓ | |

# 2. Content growth via recommendations



# How complete is Wikipedia?

Geotagged articles across all Wikipedia languages (2 million)
https://iccl.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en




Geotagged articles in English Wikipedia (950,000)

# Spanish Wikipedia

Geotagged articles in Spanish Wikipedia (261,000)

# Portuguese Wikipedia

Geotagged articles in Portuguese Wikipedia (185,000)

# Arabic Wikipedia

Geotagged articles in Arabic Wikipedia (87,000)

Depending on the language you speak, you can only learn about a small set of  topics in Wikipedia.

*How to fill these content gaps?*

http://blog.wikimedia.org/2016/04/27/article-recommendation-system/

# Article recommendations

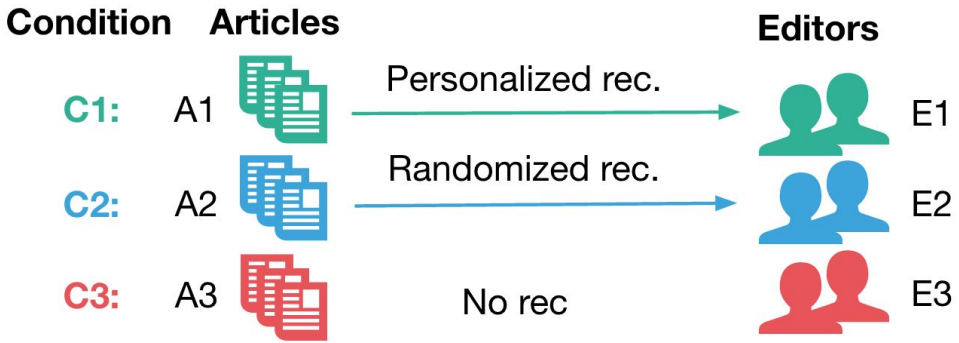
**Goal:** Recommend important missing articles to editors


In collaboration with: Stanford University

http://blog.wikimedia.org/2016/04/27/article-recommendation-system/

# Article recommendations

**Experimental results:** article creation recommendations can triple article creation rate, while controlling for quality




Wulczyn et al 2016. Growing Wikipedia Across Languages via Recommendation. *WWW '16*
http://dx.doi.org/10.1145/2872427.2883077

Wikipédia **GapFinder** beta

português ⌄ | العربية ⌄

Rio de Janeiro


Polícia Militar do Estad...

2k visualizações recentes

Volta Redonda

5k visualizações recentes

Resende (Rio de Janei...

município localizado no sul do estado do Rio de Janeiro, no Brasil

4k visualizações recentes

Geografia do Rio de Ja...

2k visualizações recentes

Petrópolis

6k visualizações recentes

Campeonato Carioca d..

6k visualizações recentes

Carioca

2k visualizações recentes

Guanabara

3k visualizações recentes

http://recommend.wmflabs.org - http://recommend.wmflabs.org/api/?t=ar&s=pt

# Understanding editing culture

WIKIMEDIA
FOUNDATION

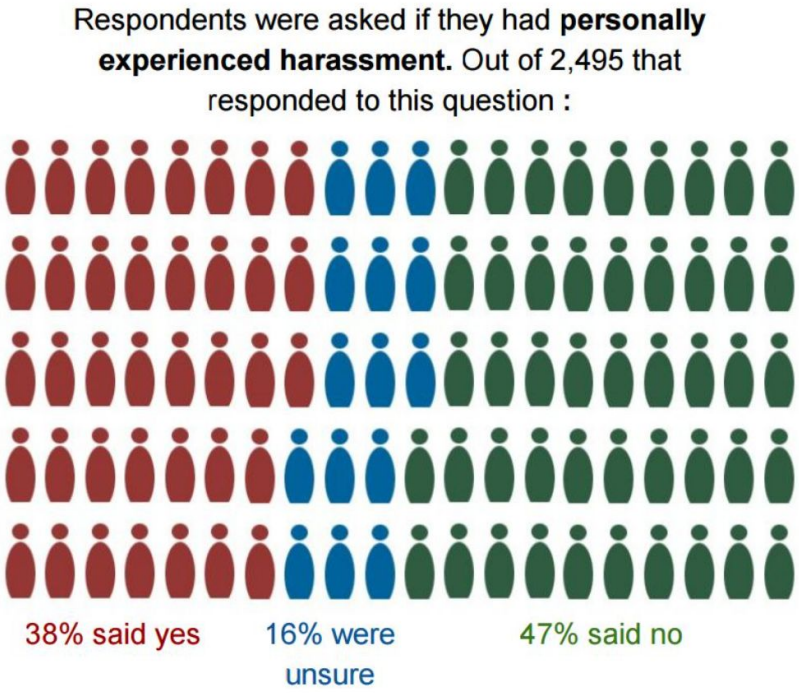
1. Toxic comments and personal attacks

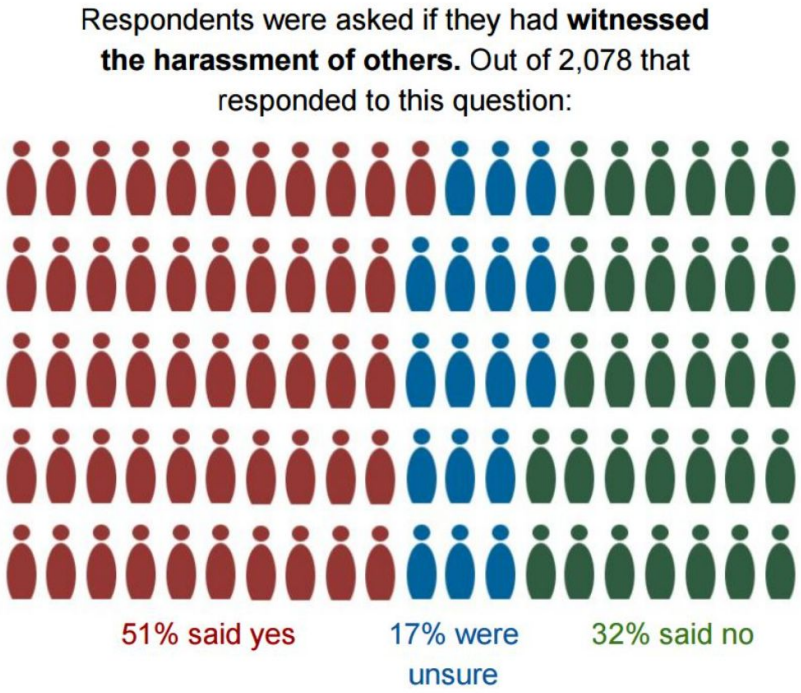
If quality control has caused new good-faith contributors to leave at increasingly higher rates

*what about toxic comments and personal attacks?*

https://meta.wikimedia.org/wiki/Research:Detox

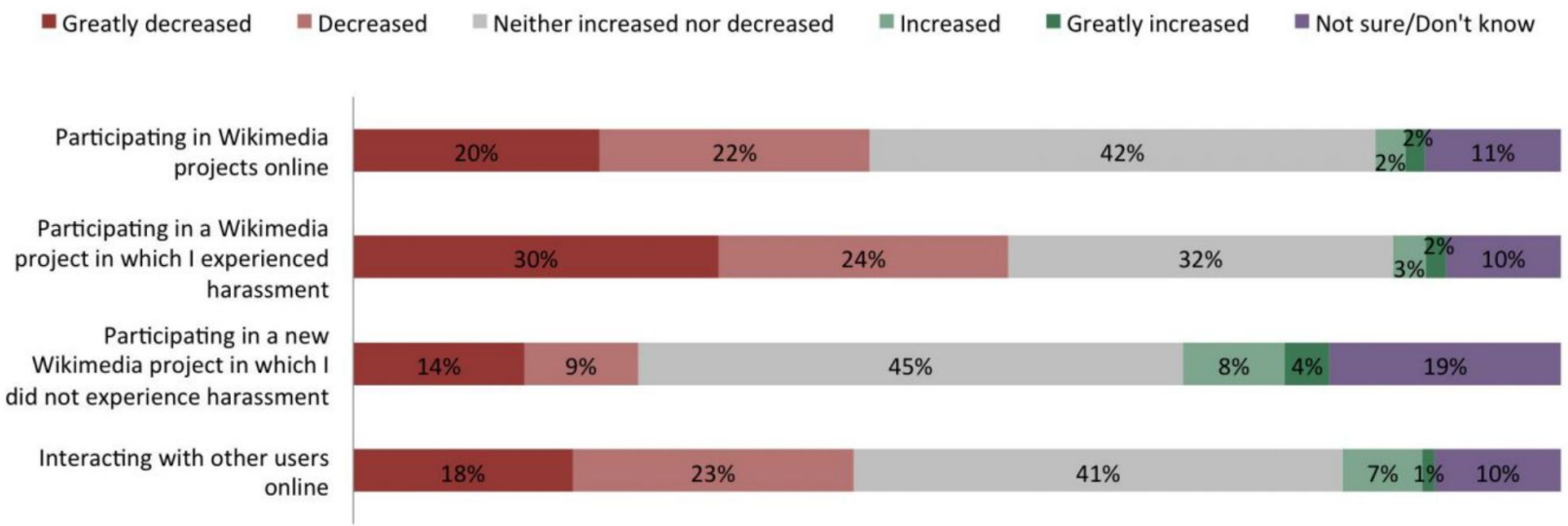
# Harassment is common in Wikipedia




https://meta.wikimedia.org/wiki/Research:Harassment_survey_2015

# Victims of harassment are less likely to contribute




https://meta.wikimedia.org/wiki/Research:Harassment_survey_2015

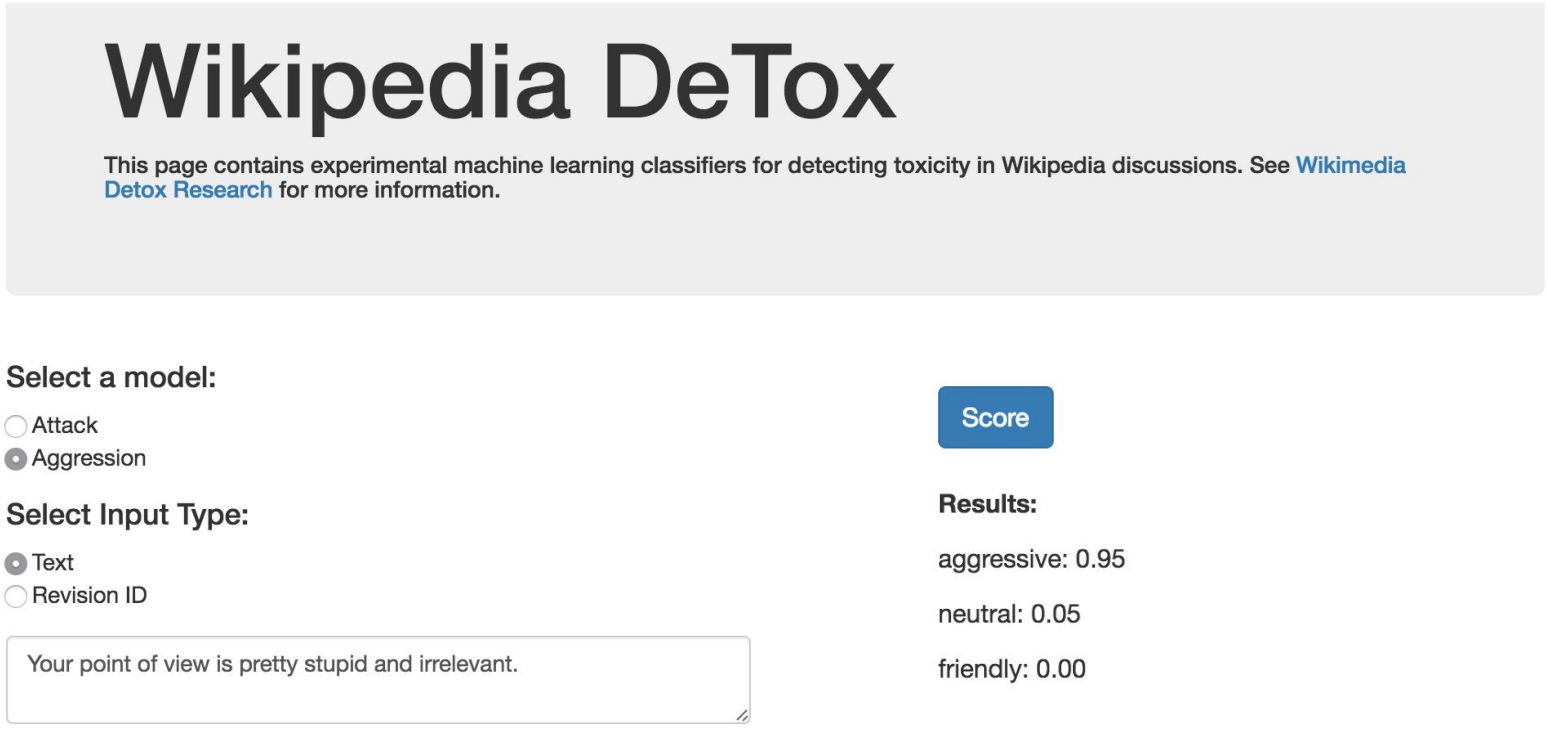
# Research:Detox

**Goals:**

1. Design algorithms to detect personal attacks on Wikipedia
2. Use these algorithms to analyze and identify harassing behavior
3. Analyze the impact of harassment on editor retention

In collaboration with: 


https://meta.wikimedia.org/wiki/Research:Detox

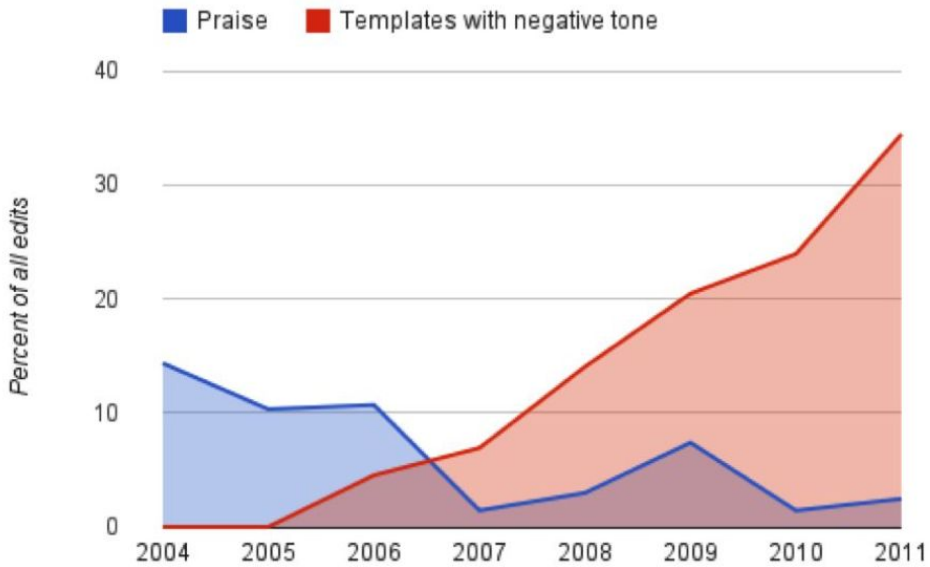



http://wikidetox.appspot.com/

# 2. Socializing newcomers

# Wikipedia is hard for new editors.

*Does a more welcoming experience increase new editor retention?*

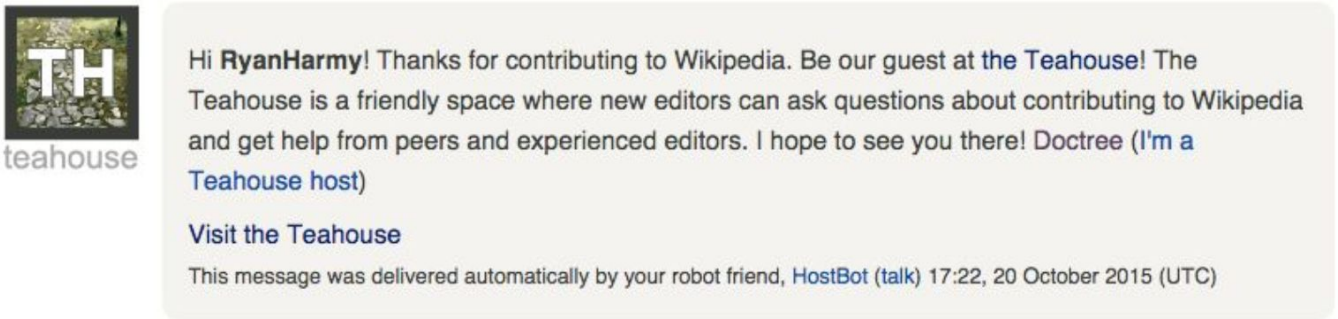



https://blog.wikimedia.org/2011/05/02/neweditorwarnings/

# Teahouse: newcomer socialization experiments

Short term 1+ edit survival: +13%

Long term 5+ edit survival: +16%




https://meta.wikimedia.org/wiki/Research:Teahouse_long_term_new_editor_retention

# Understanding Wikimedia readers

WIKIMEDIA
FOUNDATION

# 1. A taxonomy of Wikipedia readers

# Characterizing readers

*Daily pageviews*: 250 million (EN), 13 million (PT)

*Monthly unique devices*: 602 million (EN), 45 million (PT)

*Active editors*: 30,000 (EN), 1,500 (PT)

(analytics.wikimedia.org  stats.wikimedia.org, September 2016)

*Readers are a very large part of Wikipedia's ecosystem that are often overlooked*

# Characterizing readers

Webrequest logs

- IP address, user agent, page requested, etc.
- A wealth of data
- We can’t learn “why” readers read Wikipedia from webrequest logs alone

Surveys

- Surveys help understand participant motivation
- Can’t be meaningfully compared with quantitative data

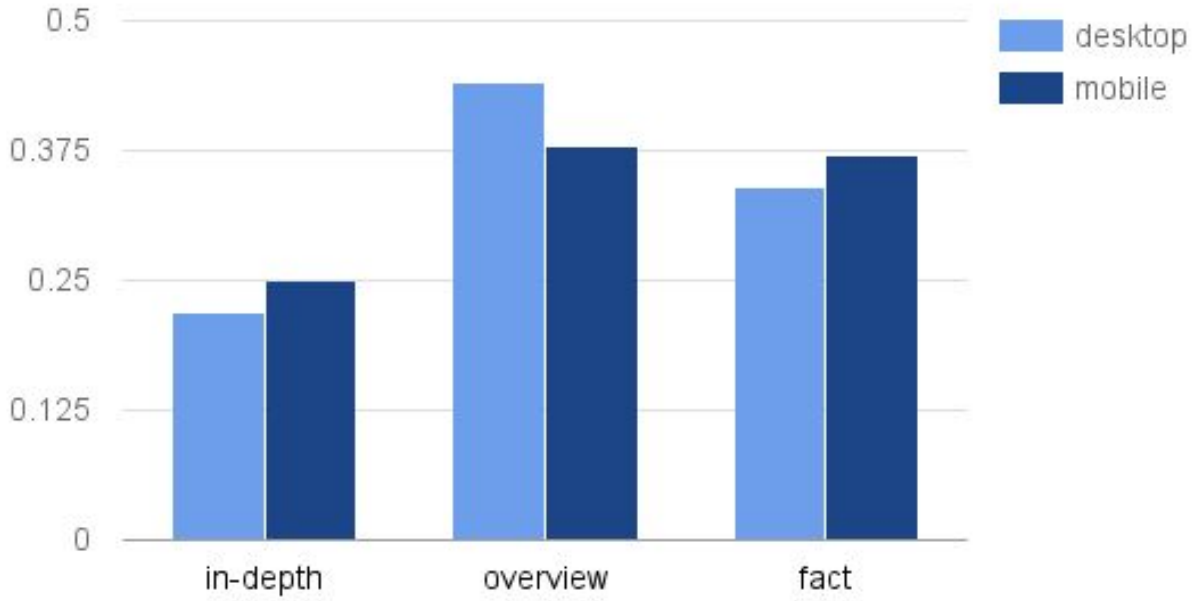
*Combine survey and webrequest data*

# A large scale reader survey

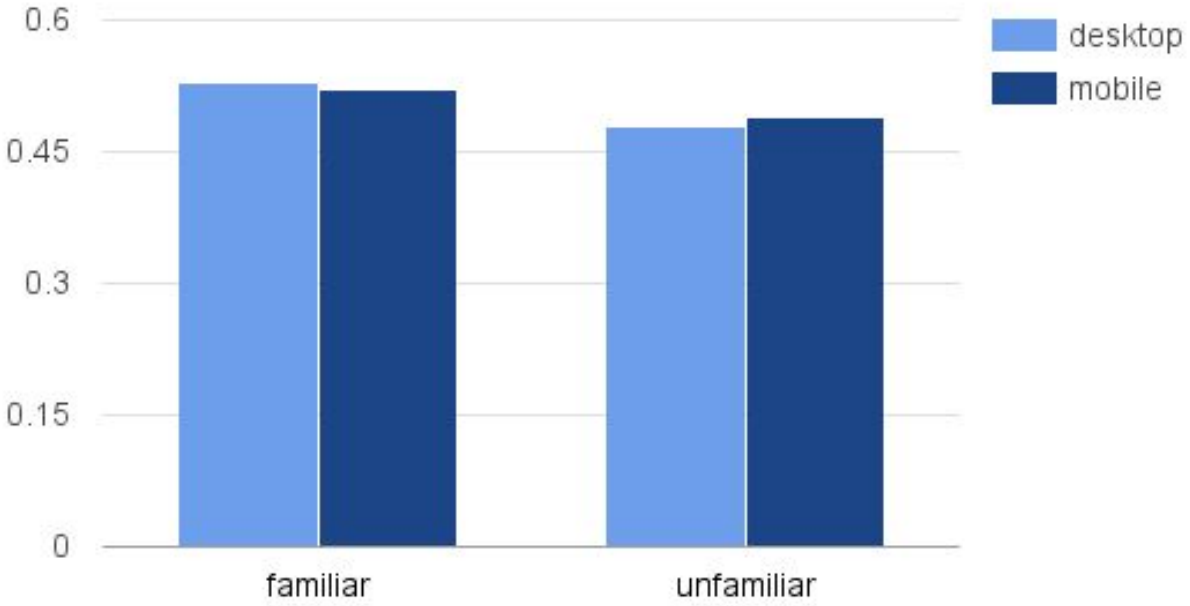
**Goal**:  Collect data on three dimensions of reader experience, link responses to webrequest log data via opt-in tokens

**Questions**: Depth of knowledge / Familiarity / Motivation
**Data:** 36K responses collected from English Wikipedia readers (Desktop+Mobile)

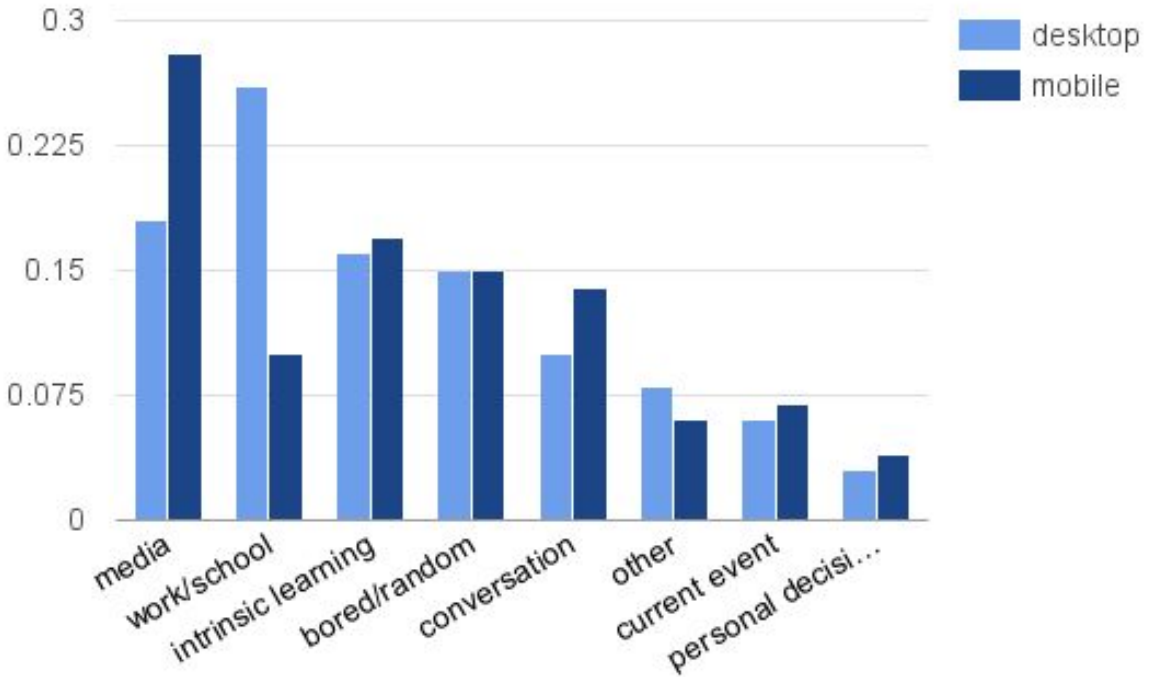
# Information Depth



# Prior knowledge



# Motivation


0.3
0.225
0.15
0.075
0
desktop
mobile
media
work/school
intrinsic learning
bored/random
conversation
other
current event
personal decisi…

# 2. New readers

# New reader research

## Countries

Mexico

India

Nigeria

## Themes

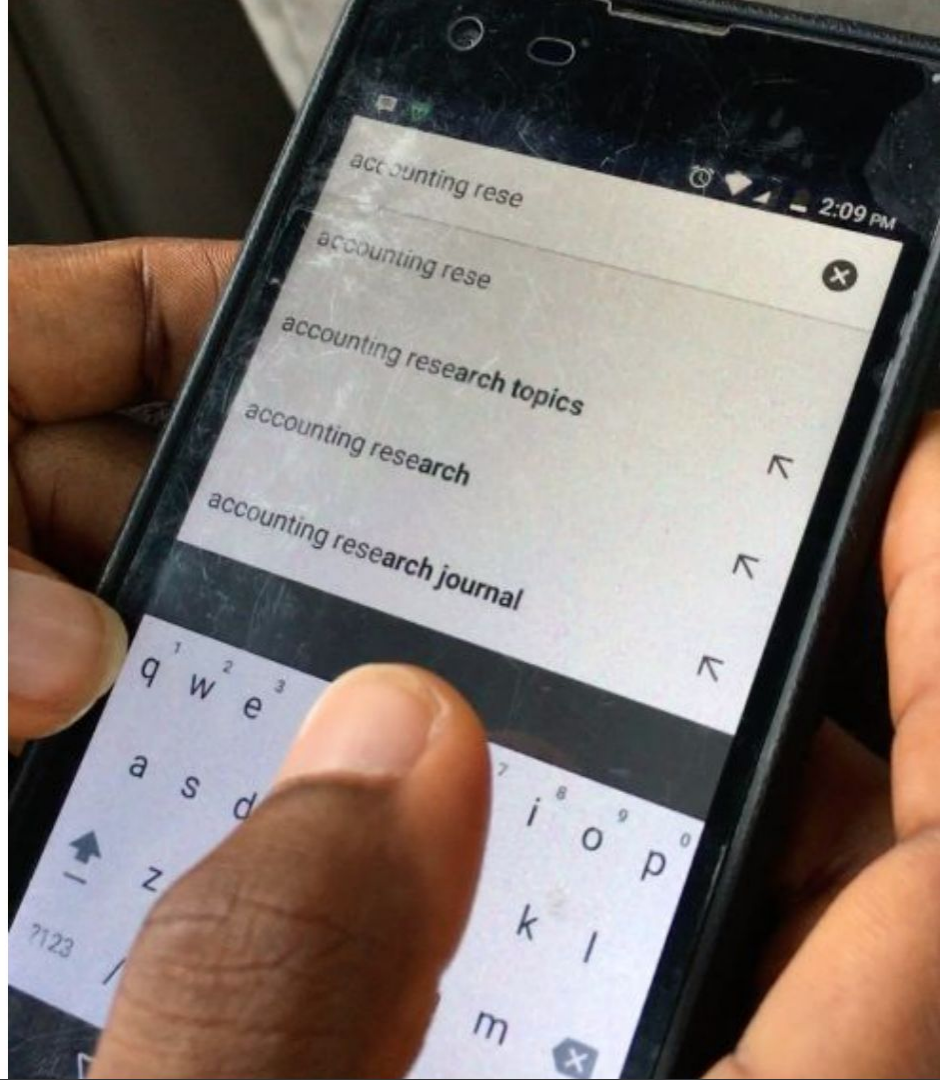
1. Information seeking
2. Accessing the internet
3. Understanding the internet
4. Using the internet
5. Getting information online
6. Wikipedia Awareness
7. Wikipedia usage

https://meta.wikimedia.org/wiki/New_Readers

## 22. People confuse Wikipedia with a search engine or social media platform. This can create unrealistic expectations of its functionality.

Learn more: Research deck, slide 73

WIKIMEDIA FOUNDATION



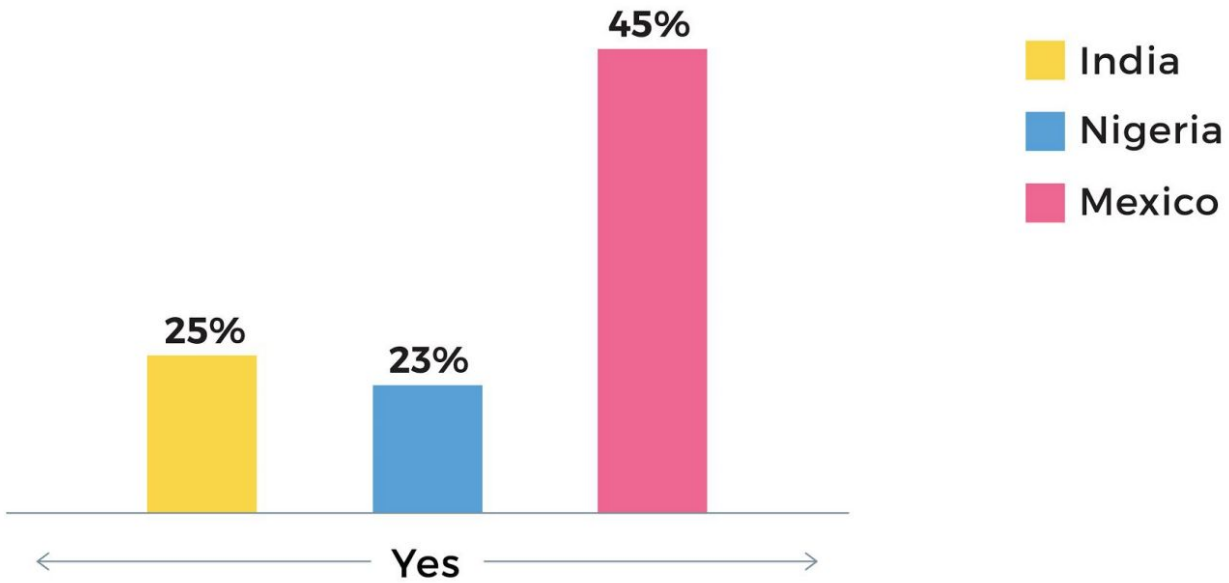
**“Wikipedia is a 'poor cousin' of Google. It is the lesser model.”**

**“Google and Wikipedia are similar. Google is more distributed; Wikipedia is more analytical and comprehensive.”**

**“Wikipedia is a social network. You’d use it if a friend in the US was on it and you wanted to connect with them.”**

Phone survey findings

## Have you ever heard of Wikipedia?




https://meta.wikimedia.org/wiki/New_Readers

# Wikimedia Research: Current priorities

- Artificial Intelligence as a service
  - *Socially aware* quality control
  - Content growth via recommendations
- Understanding editing culture
  - Toxic comments and personal attacks
  - Socializing newcomers
- Understanding readers
  - A taxonomy of Wikipedia readers
  - New readers

# Wikimedia Research: How to scale?

**1:100,000,000**

Current ratio of full-time Wikimedia researchers to Wikimedia monthly unique visitors (legacy comScore data)

# A platform for open, collaborative, and reproducible research

# Formal collaborations

WIKIMEDIA
FOUNDATION

*Carnegie Mellon University*
*Dresden University of Technology*
*GESIS - Leibniz Institute for the Social Sciences*
*Institute for Scientific Interchange*
*Jigsaw*
*Stanford University*
*University of Minnesota*
*University of Washington*
*West Virginia University*
*Wikimedia Germany*

https://www.mediawiki.org/wiki/Wikimedia_Research/Collaborators

# Open, reusable research

# Open access policy

The Wikimedia Foundation's mission is to disseminate open knowledge effectively and globally. In keeping with this mission, the Wikimedia Foundation supports research in areas that benefit the Wikimedia community. We aim to make any work produced with our support openly available to the public and reusable on Wikimedia projects.

## 1. Expectations

Researchers will need to provide unrestricted access to and reuse of all their research output if their research receives support from the Wikimedia Foundation in the form of:

- funds;
- letter of endorsement;
- equipment, hosting, or office space;
- access to non-public data or special API privileges; or
- other support under an agreement between researchers and the Wikimedia Foundation.

https://wikimediafoundation.org/wiki/Open_access_policy

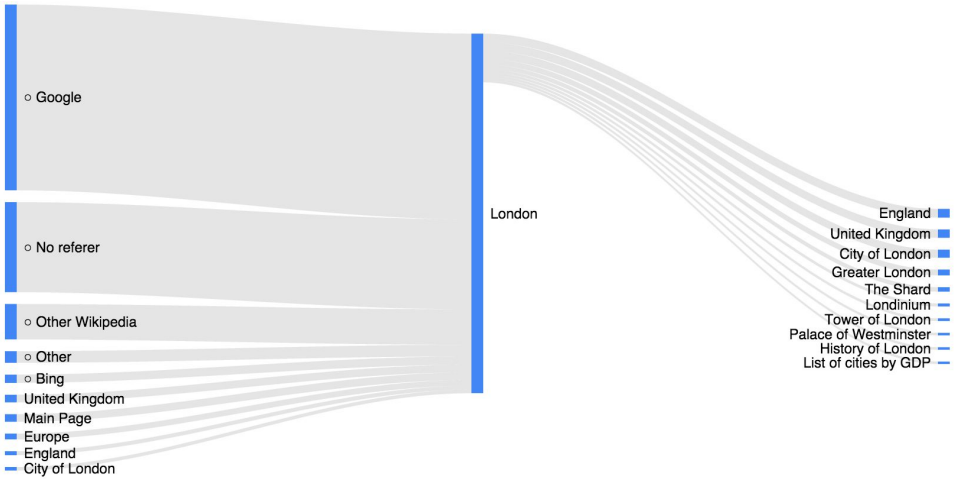
# Open data

DATA ANALYTICS, FREE KNOWLEDGE, RESEARCH, WIKIMEDIA RESEARCH NEWSLETTER, WIKIPEDIA

# *Growing free knowledge through open data*

By Dario Taraborelli, Wikimedia Foundation

March 13th, 2015



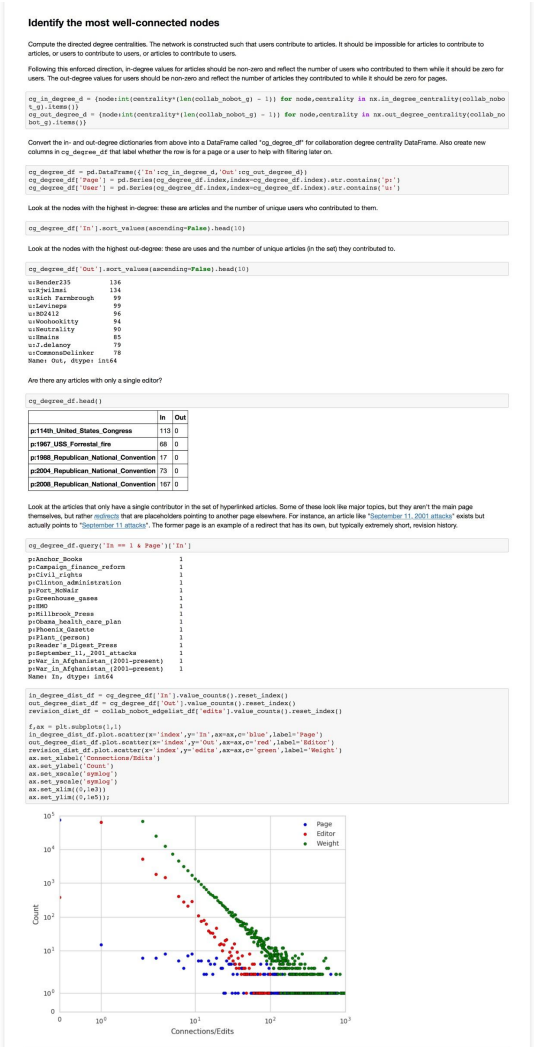

https://blog.wikimedia.org/2015/03/13/open-data-sets/

# An open notebook infrastructure

# PAWS

PAWS: A Web Shell (PAWS) hosts python notebooks and a terminal accessible with a browser.


https://paws.wmflabs.org/
https://wikitech.wikimedia.org/wiki/PAWS
Sample notebook by Brian Keegan



**Identify the most well-connected nodes**

Compute the directed degree centralities. The network is constructed such that users contribute to articles. It should be impossible for articles to contribute to articles, or users to contribute to users, or articles to contribute to users.

Following this enforced direction, in-degree values for articles should be non-zero and reflect the number of users who contributed to them while it should be zero for users. The out-degree values for users should be non-zero and reflect the number of articles they contributed to while it should be zero for pages.

```
cg_in_degree_d = {node:int(centrality*(len(collab_nobot_g) - 1)) for node,centrality in nx.in_degree_centrality(collab_nobot_g).items()}
cg_out_degree_d = {node:int(centrality*(len(collab_nobot_g) - 1)) for node,centrality in nx.out_degree_centrality(collab_nobot_g).items()}
```

Convert the in- and out-degree dictionaries from above into a DataFrame called "cg_degree_df" for collaboration degree centrality DataFrame. Also create new columns in `cg_degree_df` that label whether the row is for a page or a user to help with filtering later on.

```
cg_degree_df = pd.DataFrame({'In':cg_in_degree_d,'Out':cg_out_degree_d})
cg_degree_df['Page'] = pd.Series(cg_degree_df.index,index=cg_degree_df.index).str.contains('p:')
cg_degree_df['User'] = pd.Series(cg_degree_df.index,index=cg_degree_df.index).str.contains('u:')
```

Look at the nodes with the highest in-degree: these are articles and the number of unique users who contributed to them.

```
cg_degree_df['In'].sort_values(ascending=False).head(10)
```

Look at the nodes with the highest out-degree: these are uses and the number of unique articles (in the set) they contributed to.

```
cg_degree_df['Out'].sort_values(ascending=False).head(10)
```

```
u:Bender235          136
u:Rjwilmsi           134
u:Rich Farmbrough     99
u:ClueBot NG          99
u:BD2412              96
u:Woohookitty         94
u:Neutrality          90
u:Rmains              85
u:J.delanoy           79
u:CommonsDelinker     78
Name: Out, dtype: int64
```

Are there any articles with only a single editor?

```
cg_degree_df.head()
```

| | In | Out |
|---|---|---|
| **p:114th_United_States_Congress** | 113 | 0 |
| **p:1967_USS_Forrestal_fire** | 68 | 0 |
| **p:1988_Republican_National_Convention** | 17 | 0 |
| **p:2004_Republican_National_Convention** | 73 | 0 |
| **p:2008_Republican_National_Convention** | 167 | 0 |

Look at the articles that only have a single contributor in the set of hyperlinked articles. Some of these look like major topics, but they aren't the main page themselves, but rather redirects that are placeholders pointing to another page elsewhere. For instance, an article like "September 11, 2001 attacks" exists but actually points to "September 11 attacks". The former page is an example of a redirect that has its own, but typically extremely short, revision history.

```
cg_degree_df.query('In == 1 & Page')['In']
```

```
p:Anchor_Books                      1
p:Campaign_finance_reform           1
p:Civil_rights                      1
p:Clinton_Administration            1
p:Fort_McNair                       1
p:Greenhouse_gases                  1
p:IMD                               1
p:Millbrook_Press                   1
p:Obama_health_care_plan            1
p:Phoenix_Gazette                   1
p:Plant_(person)                    1
p:Reader's_Digest_Press             1
p:September_11,_2001_attacks        1
p:War_in_Afghanistan_(2001–present)  1
p:War_in_Afghanistan_(2001–present)  1
Name: In, dtype: int64
```

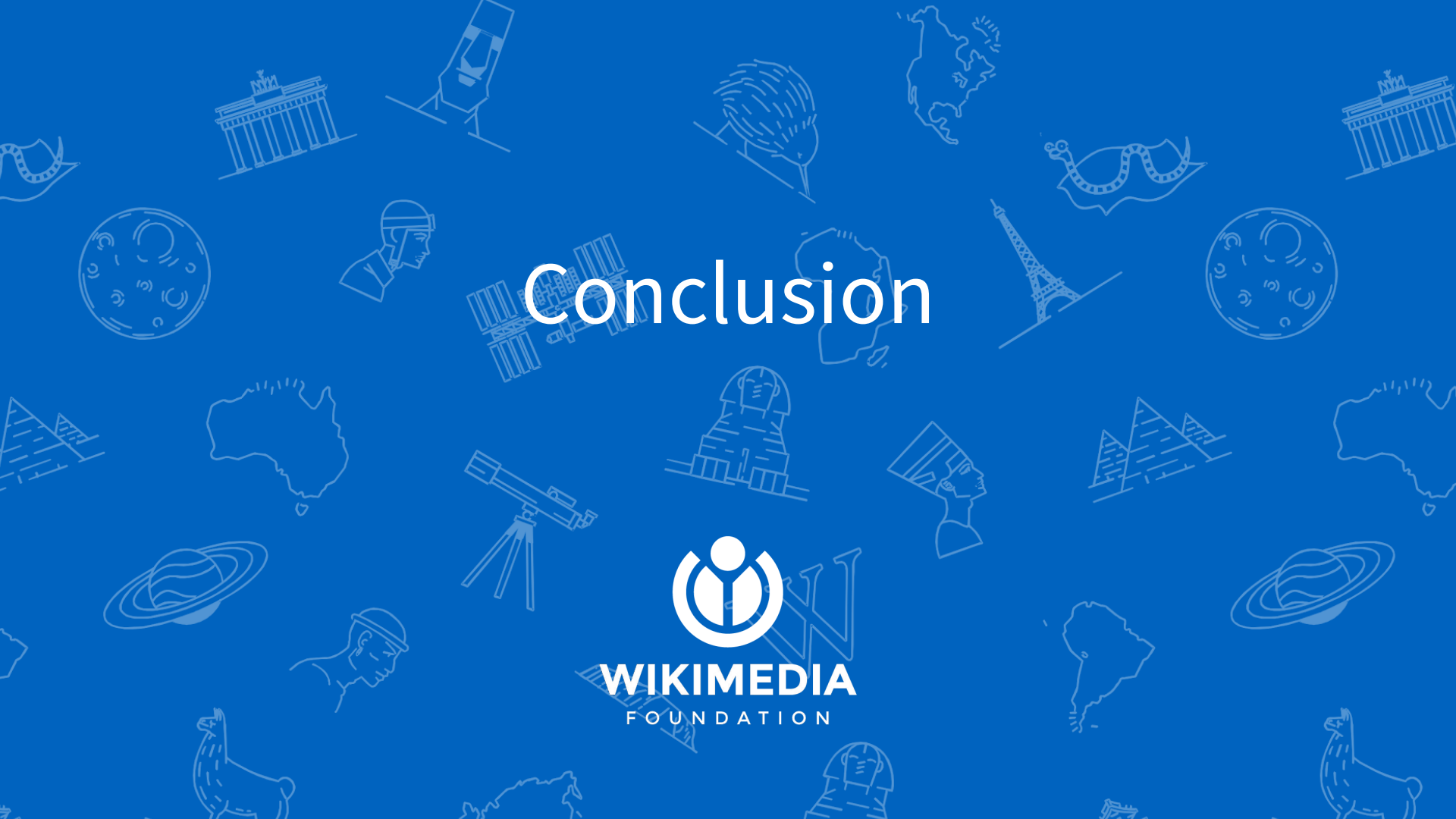
```
in_degree_dist_df = cg_degree_df['In'].value_counts().reset_index()
out_degree_dist_df = cg_degree_df['Out'].value_counts().reset_index()
revision_dist_df = collab_nobot_edgelist_df['edits'].value_counts().reset_index()

f,ax = plt.subplots(1,1)
in_degree_dist_df.plot.scatter(x='index',y='In',ax=ax,c='blue',label='Page')
out_degree_dist_df.plot.scatter(x='index',y='Out',ax=ax,c='red',label='Editor')
revision_dist_df.plot.scatter(x='index',y='edits',ax=ax,c='green',label='Weight')
ax.set_xlabel('Connections/Edits')
ax.set_ylabel('Count')
ax.set_xscale('symlog')
ax.set_yscale('symlog')
ax.set_xlim((0,1e3))
ax.set_ylim((0,1e5));
```

*learn more about*

# Open data, APIs and research tools

workshop at 2pm, auditorium PPGI

# Conclusion



[m:Research:Newsletter](#)

[@WikiResearch](#)

Wikimedia Research Newsletter

Vol: 6 • Issue: 05 • May 2016 [CONTRIBUTE] [ARCHIVES]

*English as Wikipedia's lingua franca; deletion rationales; schizophrenia controversies*

**With contributions by:** Morten Warncke-Wang, Piotr Konieczny, Federico Leva, Steve Jankowski and Tilman Bayer

## English still the lingua franca of Wikipedia

*Reviewed by Morten Warncke-Wang*

Many of the more active Wikipedia contributors are multilingual. In the April 2011 Wikipedia Editors Survey,[supp 1] 72% of respondents said they read Wikipedia content in more than one language, and 51% said they contributed to multiple Wikipedias. Research has estimated that approximately 15% of active Wikipedians are multilingual.[supp 2] These contributors are important as they can enable knowledge transfer between different language editions of Wikipedia, yet little is known about who they are and what they do.




Primary and non-primary editors show different engagement levels (figure 4 from the paper)

A recent paper published in PLOS ONE by researchers at KAIST and OII, titled "Understanding Editing Behaviors in Multilingual Wikipedia"[1], adds to our knowledge of multilingual contributors by investigating their engagement level, topic interests, and language proficiency. The paper uses a dataset spanning a month of Wikipedia contributions in July–August 2013 and defines a *multilingual editor* as one who make contributions to multiple languages. Overall the dataset contains 12,577

# Thank you

**Acknowledgments**

*The Wikimedia Research team, our fellows and collaborators:*
Samantha Becker, Grage Gellerman, Aaron Halfaker, Jonathan Morgan, Yuvi Panda, Abbey Ripstra, Amir Sarabadani, Nathaniel Schaaf, Morten Warncke-Wang, Robert West, Erik Zachte, Leila Zia

*Our formal collaborations:*
mw:Wikimedia_Research/Collaborators

dario@wikimedia.org • @readermeter • @WikiResearch