

Ethical AI, Challenges, and

a Pitch

Beyond transparency



The Free Encyclopedia

4 979 000+ articles

Deutsch

Die freie Enzyklopädie

1 861 000+ Artikel

Русский

Свободная энциклопедия

1 257 000+ статей

中文

自由的百科全书

843 000+ 條目

Português

A enciclopédia livre

890 000+ artigos

La enciclopedia libre

1 205 000+ artículos

フリー百科事典

985 000+ 記事

Français

L'encyclopédie libre

1 205 000+ articles

Italiano

L'enciclopedia libera

1 226 000+ voci

Polski

Wolna encyklopedia

1 136 000+ haseł

Aaron Halfaker
ahalfaker@wikimedia.org

Aaron Halfaker

Principal Research Scientist, Wikimedia Foundation

Think big. Measure what you can. Build better technologies.



About me

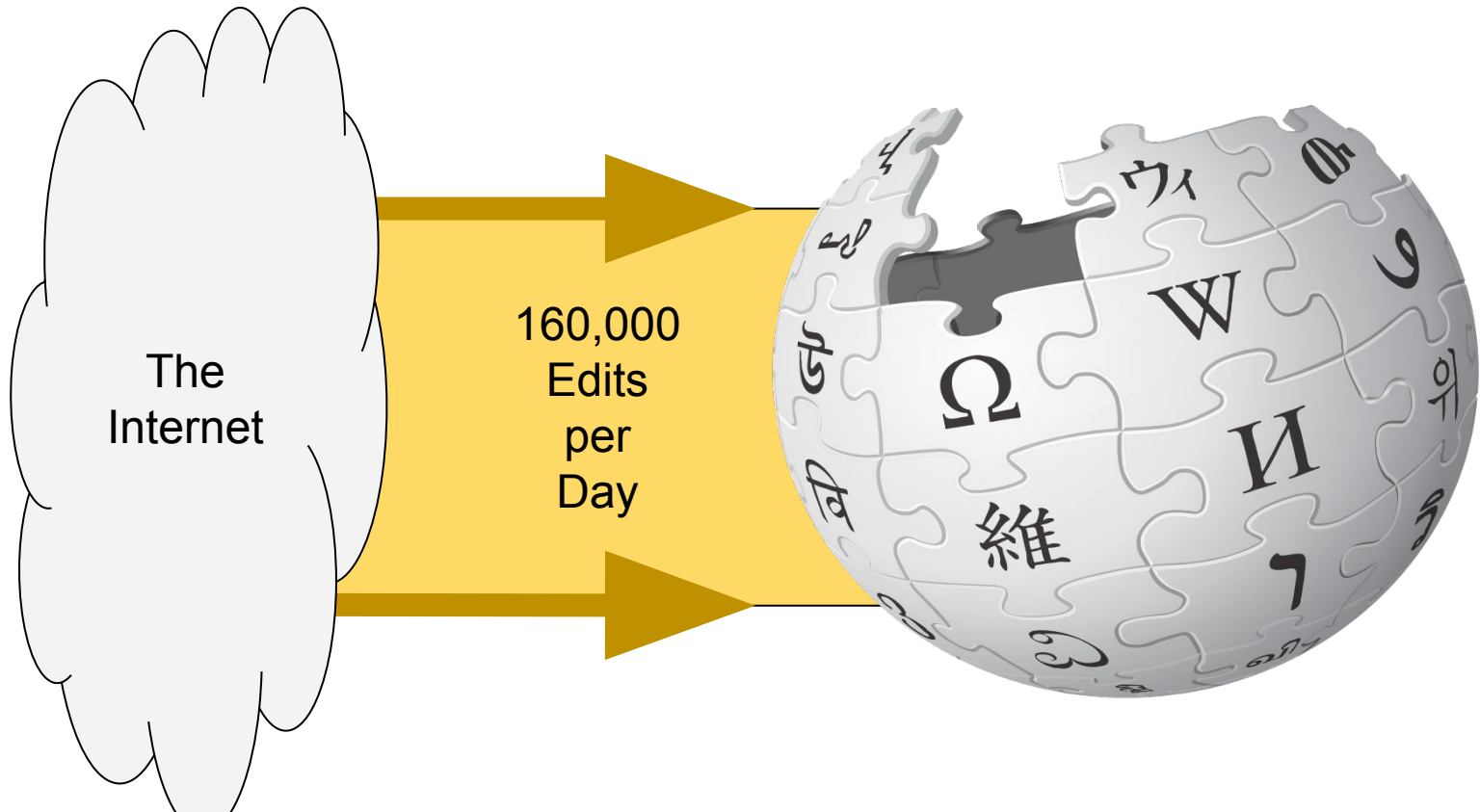
Hi. I'm Aaron Halfaker. I'm a scientist. See [projects](#) and [publications](#) below. I've been a Wikipedian since 2008. I mostly build tools

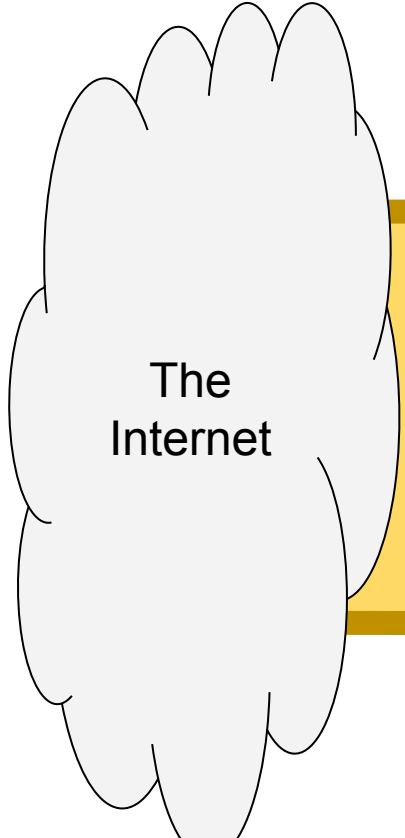
My work

My job is to build understanding about and support for the socio-technical fabric of the Wikimedia movement. I tend to focus on

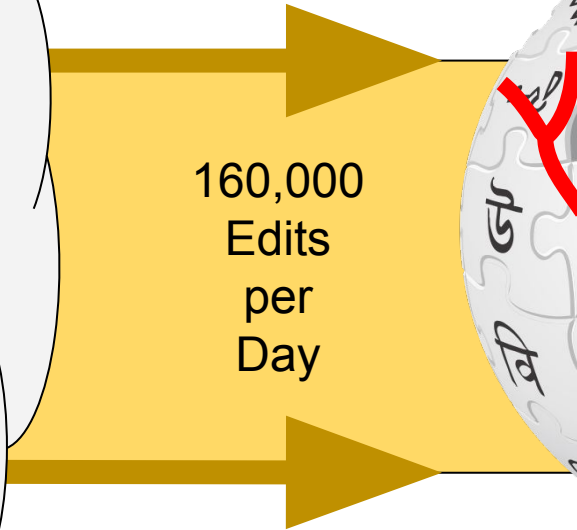
Contact me

- E-mail: ahalfaker@wikimedia.org
- Website: <http://halfaker.info> 
- Twitter: <http://twitter.com/halfaker> 





The
Internet



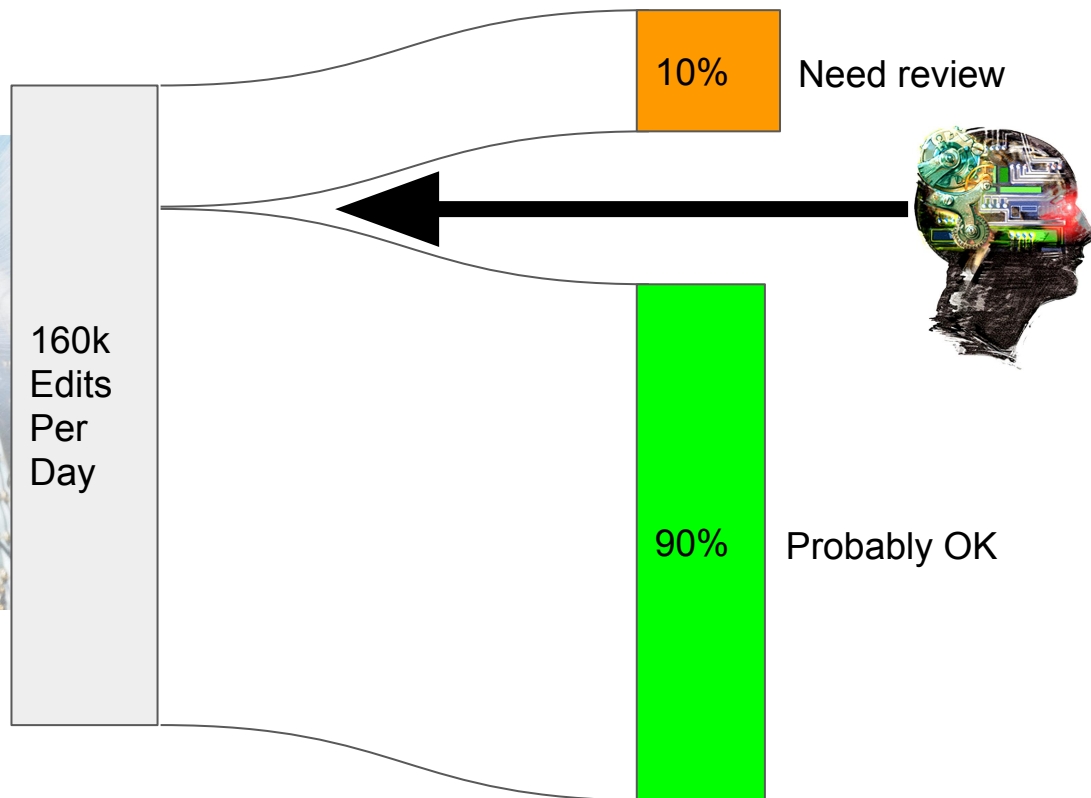
160,000
Edits
per
Day



Counter vandalism



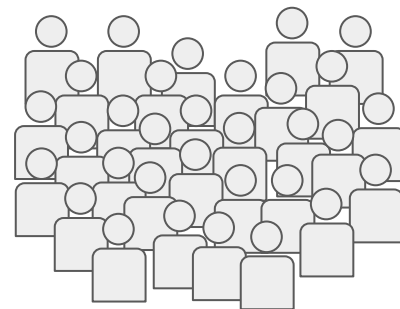
Photo taken by US Navy (PD)



Without Machine Prediction: Reviewing 160k edits per day...

~300 Hours

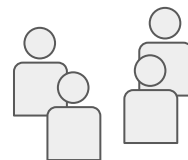
(33 people * 8 hours)



With Machine Prediction: Reviewing 16k edits per day...

~28 Hours

(4 people * 8 hours)





Machine Prediction helps
make Wikipedia better



Machine Prediction helps
make Wikipedia better

But it makes quality control
decisions more opaque.



How will our volunteer
editors maintain control?

Part 2: Challenges

<http://ores.wmflabs.org/scores/enwiki/damaging/638307884>



<http://ores.wmflabs.org/scores/enwiki/damaging/638307884>

“English Wikipedia”



<http://ores.wmflabs.org/scores/enwiki/damaging/638307884>

Is this edit damaging?



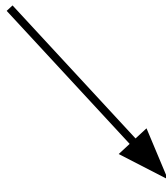
<http://ores.wmflabs.org/scores/enwiki/damaging/638307884>

<http://ores.wmflabs.org/scores/enwiki/damaging/638307884>

In 2013 Wilson, in collaboration with [[Mikhail Baryshnikov]] and co-starring [[M Dafoe]], developed "[[The Old Woman (play)|The Old Woman]]", an adaptatio work by the Russian author [[Daniil Kharms]]. The play premiered at MIF13, M: International Festival.<ref>{{Cite web|url = http://www.mif.co.uk/event/the-old-v = The Old Woman Robert Wilson, Mikhail Baryshnikov, Willem Dafoe|date = 2013|accessdate = 2014-11-23|website = Manchester International Festival|pu Manchester International Festival|last = Jansch|first = Lucie}}</ref> Wilson wro and [[Mikhail Baryshnikov|Baryshnikov]] have discussed creating a play togeth

<http://ores.wmflabs.org/scores/enwiki/damaging/638307884>

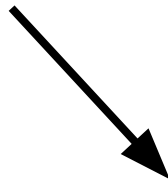
In 2013 Wilson, in collaboration with [[Mikhail Baryshnikov]] and co-starring [[M Dafoe]], developed "[[The Old Woman (play)|The Old Woman]]", an adaptatio work by the Russian author [[Daniil Kharms]]. The play premiered at MIF13, M: International Festival.<ref>{{Cite web|url = http://www.mif.co.uk/event/the-old-v = The Old Woman Robert Wilson, Mikhail Baryshnikov, Willem Dafoe|date = 2013|accessdate = 2014-11-23|website = Manchester International Festival|pu Manchester International Festival|last = Jansch|first = Lucie}}</ref> Wilson wro and [[Mikhail Baryshnikov|Baryshnikov]] have discussed creating a play togeth



```
"638307884": {  
  "prediction": false,  
  "probability": {  
    "false": 0.942,  
    "true": 0.058  
  }  
}
```


<http://ores.wmflabs.org/scores/enwiki/damaging/638307884>

In 2013 Wilson, in collaboration with [[Mikhail Baryshnikov]] and co-starring [[M Dafoe]], developed "[[The Old Woman (play)|The Old Woman]]", an adaptatio work by the Russian author [[Daniil Kharms]]. The play premiered at MIF13, M: International Festival.<ref>{{Cite web|url = http://www.mif.co.uk/event/the-old-v = The Old Woman Robert Wilson, Mikhail Baryshnikov, Willem Dafoe|date = 2013|accessdate = 2014-11-23|website = Manchester International Festival|pu Manchester International Festival|last = Jansch|first = Lucie}}</ref> Wilson wro and [[Mikhail Baryshnikov|Baryshnikov]] have discussed creating a play togeth



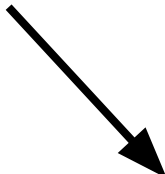
```
"638307884": {  
  "prediction": false,  
  "probability": {  
    "false": 0.942,  
    "true": 0.058  
  }  
}
```

<http://ores.wmflabs.org/scores/enwiki/damaging/642215410>

+ 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**

<http://ores.wmflabs.org/scores/enwiki/damaging/638307884>

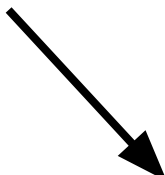
In 2013 Wilson, in collaboration with [[Mikhail Baryshnikov]] and co-starring [[M Dafoe]], developed "[[The Old Woman (play)|The Old Woman]]", an adaptatio work by the Russian author [[Daniil Kharms]]. The play premiered at MIF13, M: International Festival.<ref>{{Cite web|url = http://www.mif.co.uk/event/the-old-v = The Old Woman Robert Wilson, Mikhail Baryshnikov, Willem Dafoe|date = 2013|accessdate = 2014-11-23|website = Manchester International Festival|pu Manchester International Festival|last = Jansch|first = Lucie}}</ref> Wilson wro and [[Mikhail Baryshnikov|Baryshnikov]] have discussed creating a play togeth



```
"638307884": {  
  "prediction": false,  
  "probability": {  
    "false": 0.942,  
    "true": 0.058  
  }  
}
```

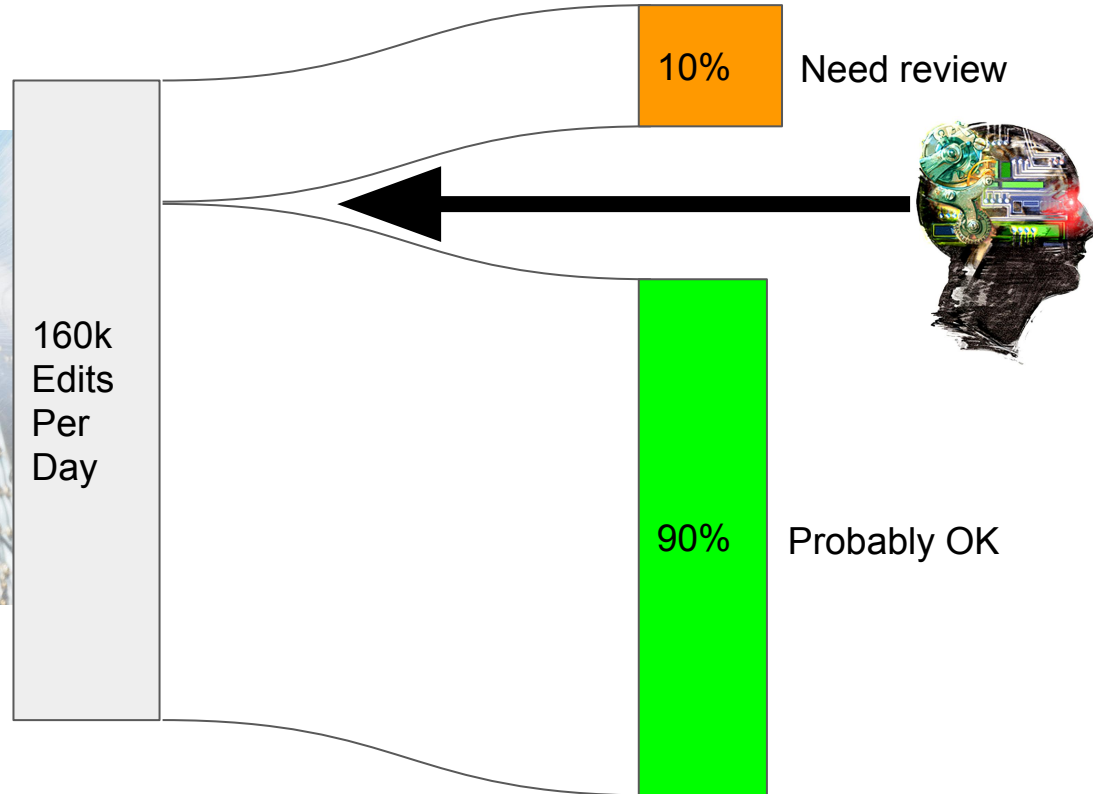
<http://ores.wmflabs.org/scores/enwiki/damaging/642215410>

+ 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": 0.080,  
    "true": 0.920  
  }  
}
```

Counter vandalism



Anonymous editors

Anonymous editors

📍 Maniphest > T118982

✓ **hewiki "reverted" model weights strongly against anons**

✓ Closed, Resolved

🌐 Public

📍 Maniphest > T129624

✓ **Investigate nlwiki 'reverted' model seems broken (always ~0.89 for anonymous edits)**

✓ Closed, Resolved

🌐 Public

Otherwise, anons seemed to dominate false-positive reports from every wiki

**... maybe anons are really
bad.**

... **maybe anons are really bad.**

- Generally, anon edits are **twice** as likely to be vandalism

... maybe anons are really bad.

- Generally, anon edits are **twice** as likely to be vandalism
- **90%** of anonymous edits are good

https://ores.wmflabs.org/v2/scores/enwiki/damaging/642345235?feature.revision.user.is_anon=false

https://ores.wmflabs.org/v2/scores/enwiki/damaging/642345235?feature.revision.user.is_anon=false

```
{"prediction": false,  
  "probability": {"false": 0.656,  
                 "true": 0.344}}
```

https://ores.wmflabs.org/v2/scores/enwiki/damaging/642345235?feature.revision.user.is_anon=false

```
{"prediction": false,  
  "probability": {"false": 0.656,  
                 "true": 0.344}}
```

https://ores.wmflabs.org/v2/scores/enwiki/damaging/642345235?feature.revision.user.is_anon=true

```
{"prediction": false,  
  "probability": {"false": 0.541,  
                 "true": 0.459}}
```



Disparate impact

From Wikipedia, the free encyclopedia

Disparate impact in [United States labor law](#) refers to practices in employment, housing, and other areas that adversely affect one group of people of a protected characteristic more than another, even though rules applied by employers or landlords are formally neutral. Although the protected classes vary by statute, most federal civil rights laws protect based on race, color, religion, national origin, and sex as protected traits, and some laws include disability status and other traits as well.

A violation of Title VII of the [1964 Civil Rights Act](#) may be proven by showing that an employment practice or policy has a disproportionately adverse effect on members of the protected class as compared with non-members of the protected class.^[1] Therefore, the disparate impact theory under Title VII prohibits employers "from using a facially neutral employment practice that has an unjustified adverse impact on members of a protected class. A facially neutral employment practice is one that does not appear to be discriminatory on its face; rather it is one that is discriminatory in its application or effect."^[2] Where a disparate impact is shown, the plaintiff can prevail without the necessity of showing intentional discrimination unless the defendant employer demonstrates that the practice or policy in question has a demonstrable relationship to the requirements of the job in question.^[3] This is the "business necessity" defense.^[1]

In addition to Title VII, other federal laws also have disparate impact provisions, including the [Age Discrimination in Employment Act of 1967](#).^[4] Some civil rights laws, such as [Title VI of the Civil Rights Act of 1964](#), do not contain disparate impact provisions creating a private right of action,^[5] although the federal government may still pursue disparate impact claims under these laws.^[6] The U.S. Supreme Court has held that the [Fair Housing Act of 1968](#) creates a cause of action for disparate impact.^[7] Disparate impact contrasts with [disparate treatment](#). A disparate impact is unintentional, whereas a disparate treatment is an intentional decision to treat people differently based on

Disparate impact

From Wikipedia, the free encyclopedia

Dis
peo
the
pro
A vi
adv
the
me
one
sho
rela

... practices in employment, housing, and other areas that adversely affect one group of people of a protected characteristic more than another ...

roup of
hough
onately
e impact
on
ther it is
sity of
trable

In addition to Title VII, other federal laws also have disparate impact provisions, including the [Age Discrimination in Employment Act of 1967](#).^[4] Some civil rights laws, such as [Title VI of the Civil Rights Act of 1964](#), do not contain disparate impact provisions creating a private right of action,^[5] although the federal government may still pursue disparate impact claims under these laws.^[6] The U.S. Supreme Court has held that the [Fair Housing Act of 1968](#) creates a cause of action for disparate impact.^[7] Disparate impact contrasts with [disparate treatment](#). A disparate impact is unintentional, whereas a disparate treatment is an intentional decision to treat people differently based on

Disparate impact

From Wikipedia, the free encyclopedia

Dis

peo

the

pro

A vi

adv

the

me

one

sho

rela

In a

196

right

has

... practices in employment, housing, and other areas that adversely affect one group of people of a protected characteristic more than another ...

race, sex, or ethnic group

held that the Fair Housing Act of 1968 creates a cause of action for disparate impact. Disparate impact contrasts with [disparate treatment](#). A disparate impact is unintentional, whereas a disparate treatment is an intentional decision to treat people differently based on

group of

though

ionately

e impact

on

ther it is

sity of

trable

of

private

Court

**Should newcomers and
anonymous status be
considered “protected”?**

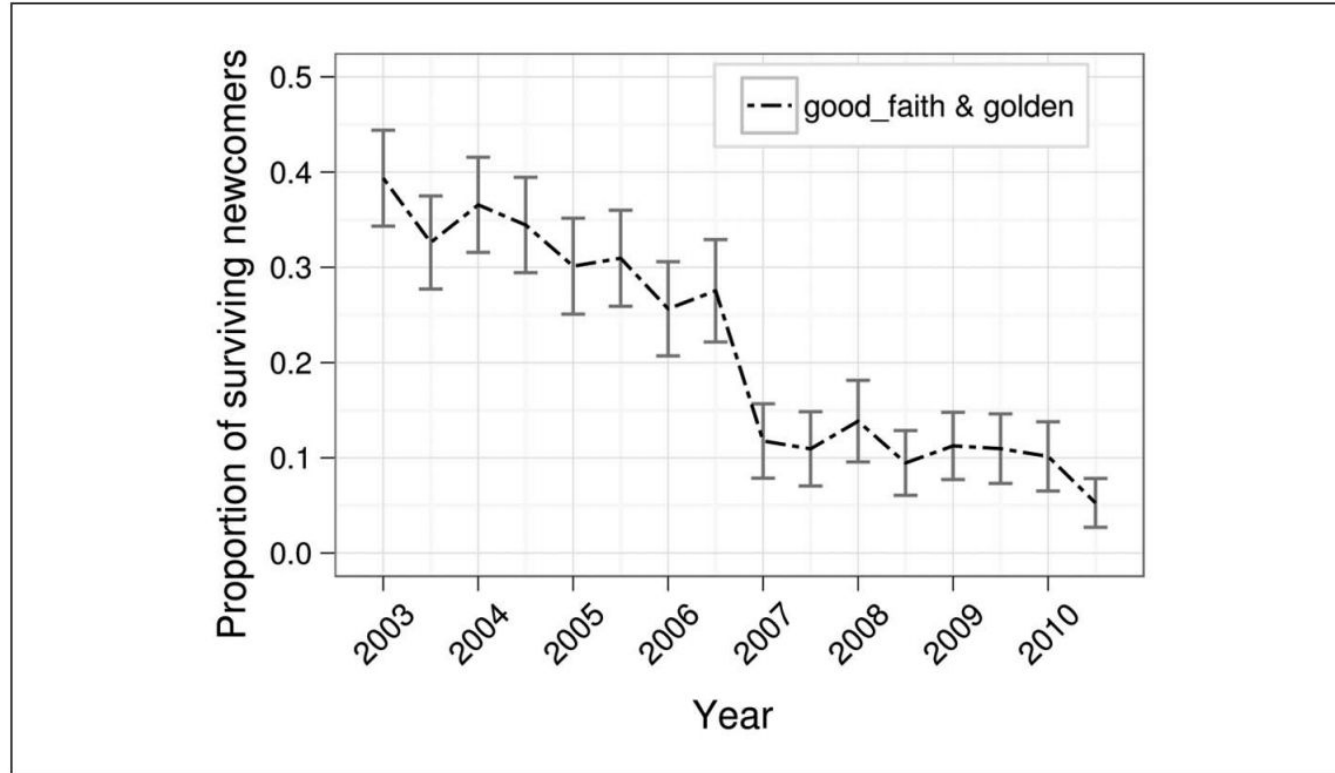


Figure 4. Survival of desirable newcomers over time.

The proportion of surviving good (“good faith” and “golden” combined) newcomers is plotted over time.

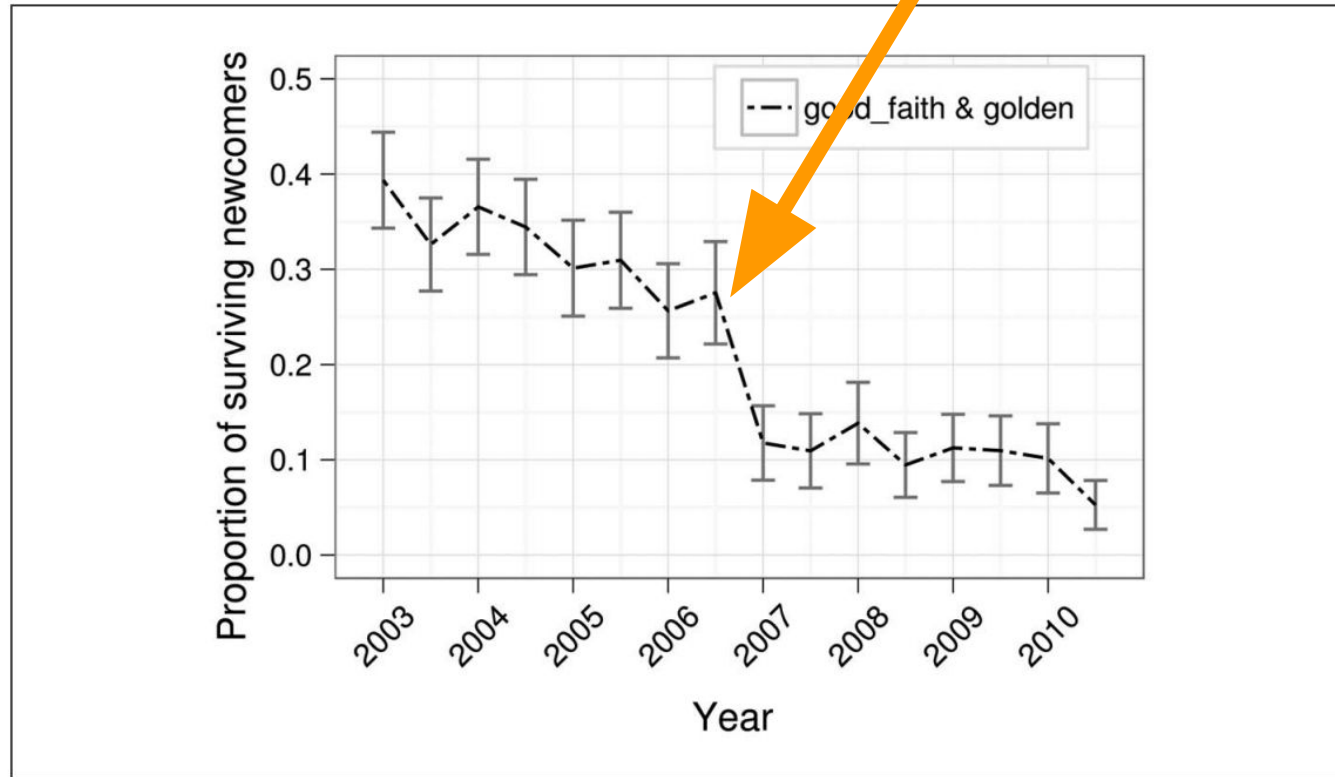


Figure 4. Survival of desirable newcomers over time.

The proportion of surviving good (“good faith” and “golden” combined) newcomers is plotted over time.

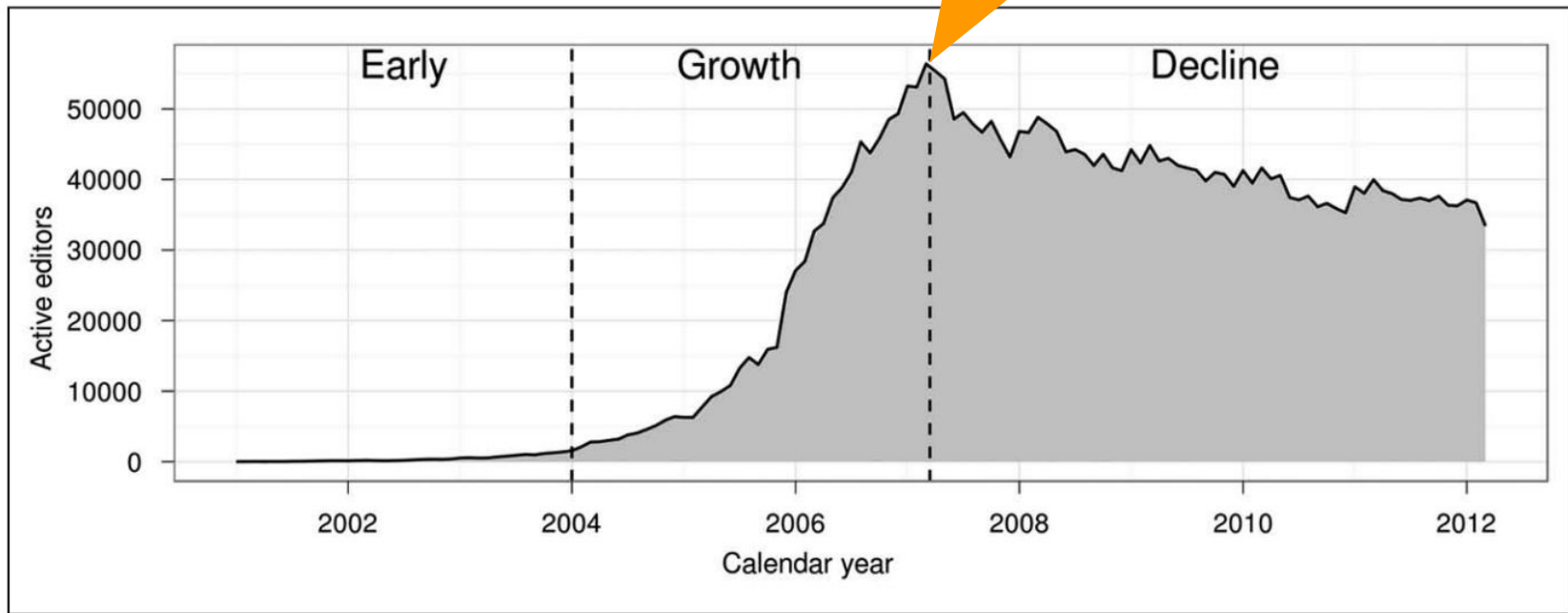


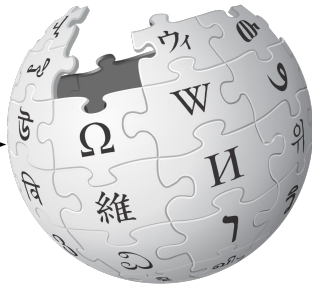
Figure 2. The English Wikipedia's editor decline.

The number of active, registered editors (≥ 5 edits per month) is plotted over time.

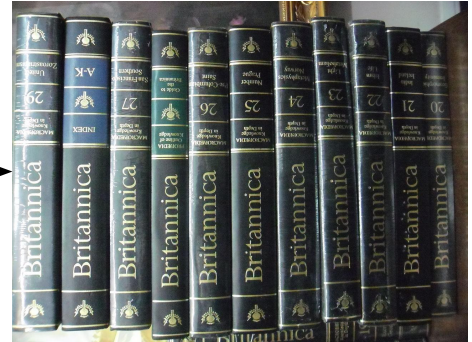
Available
Human
Attention



Input

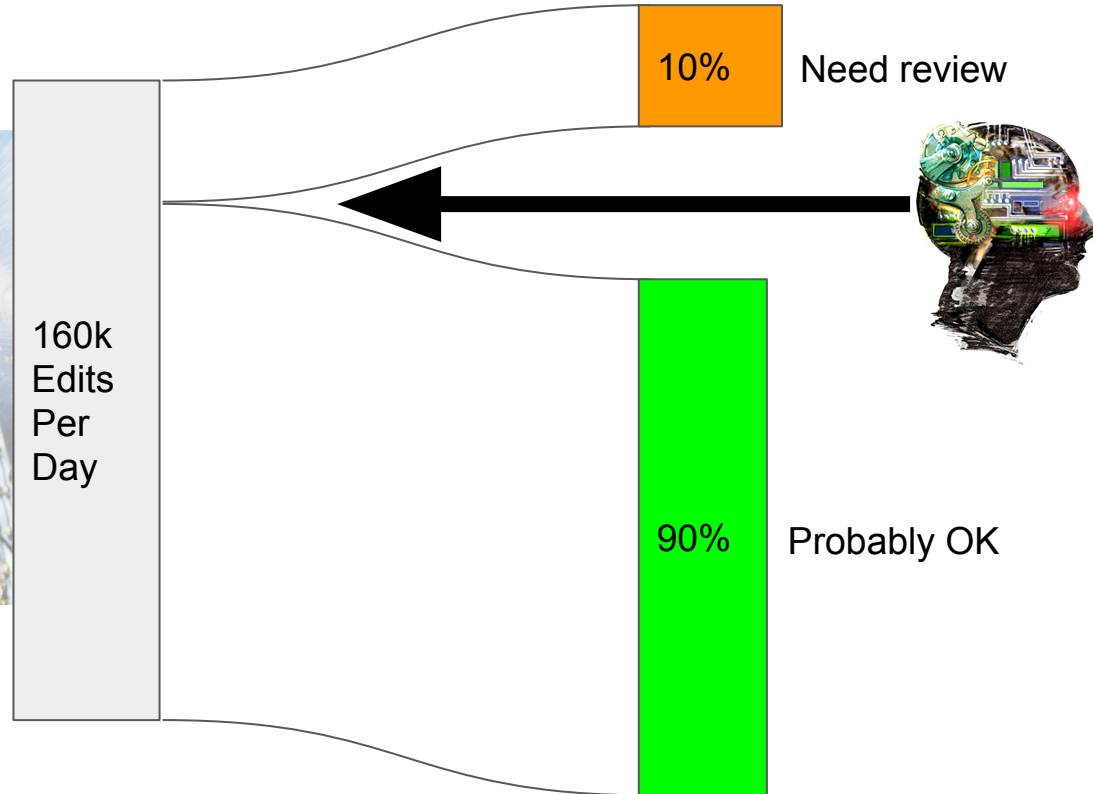


Output

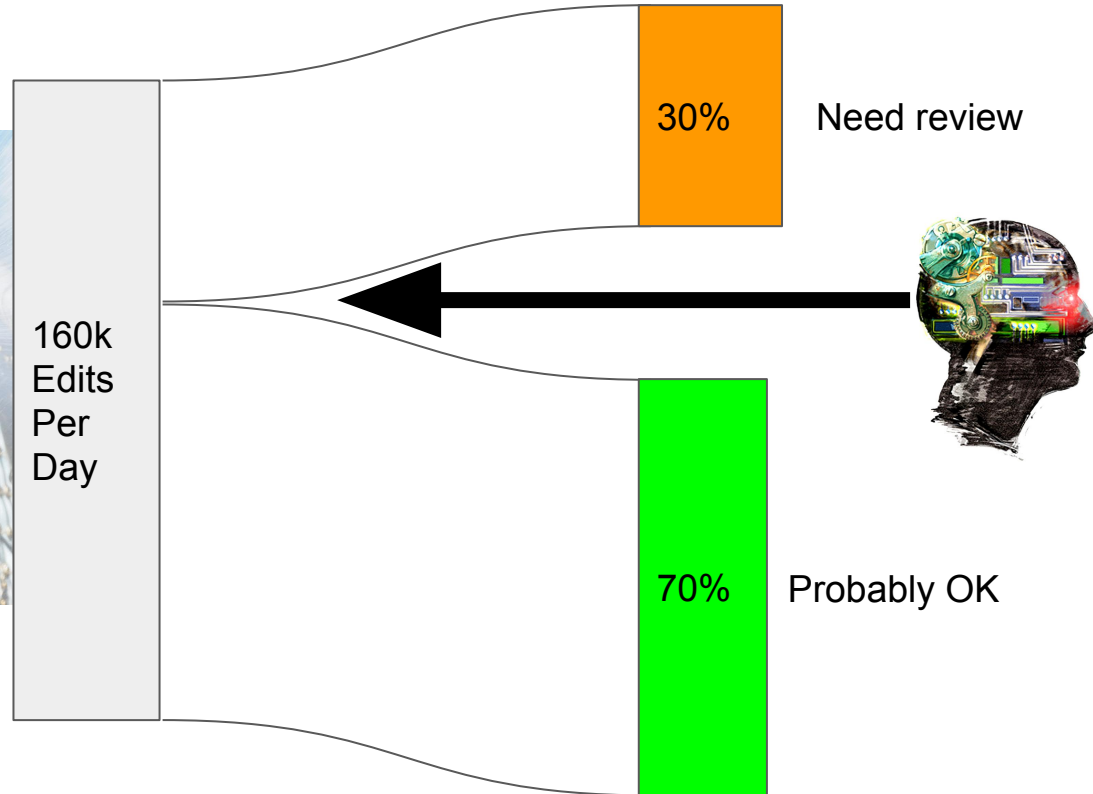


Values

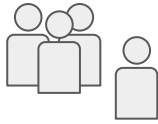
Counter vandalism w/ “user features”



Counter vandalism w/o “user features”

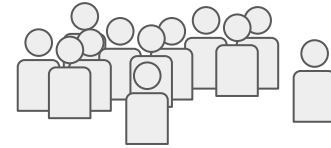


With User Features
and a strong bias against
newcomers and anonymous
editors

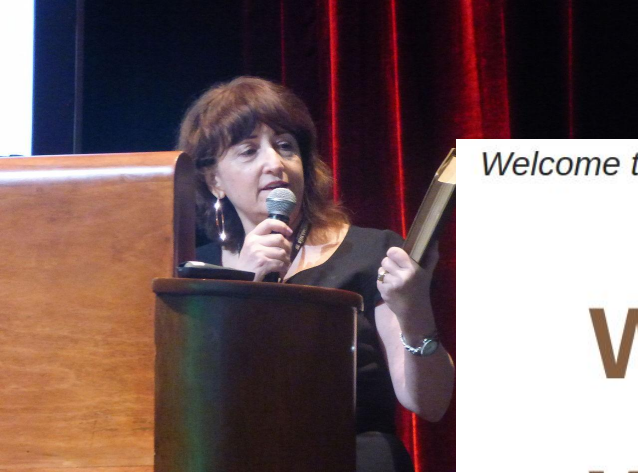


27 hours
4 people

Without User Features
More fairness
Reduced efficiency



81 hours
12 people



[Photo](#) By Luisalvaz CC-BY-SA 4.0

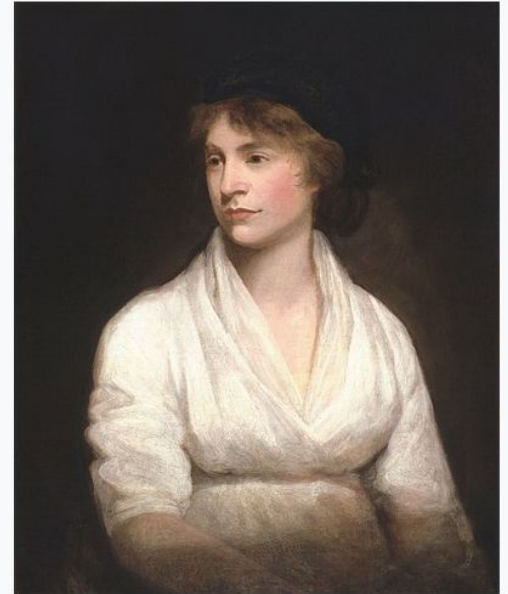
Welcome to...

WikiProject Women writers

A [WikiProject](#) dedicated to ensuring quality coverage of women writer biographies, as well as their works and their awards. The project also

[Mary Wollstonecraft by John Opie \(c. 1797\).jpg](#) (PD)

WikiProject Women writers



Mary Wollstonecraft

Shortcut

[WP:WMNWRITE](#)

Categories

[WikiProject Women writers articles](#), [Women writers](#)

Portals



[Biography portal](#)



Efficiency!

Patrollers

Fairness for newcomers!

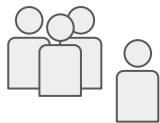


Rosiestep

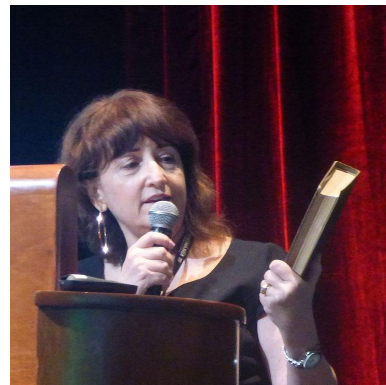
Whose values win? How do we balance?

Efficiency!

Fairness for newcomers!



Patrollers



Rosiestep



Let's teach them the basics of machine learning! Then, they'll understand what they are working with.



Let's teach them the basics of machine learning! Then, they'll understand what they are working with.

Precision Recall
Cross-validation
Feature engineering
Hyperparameter optimization
Sample weighting

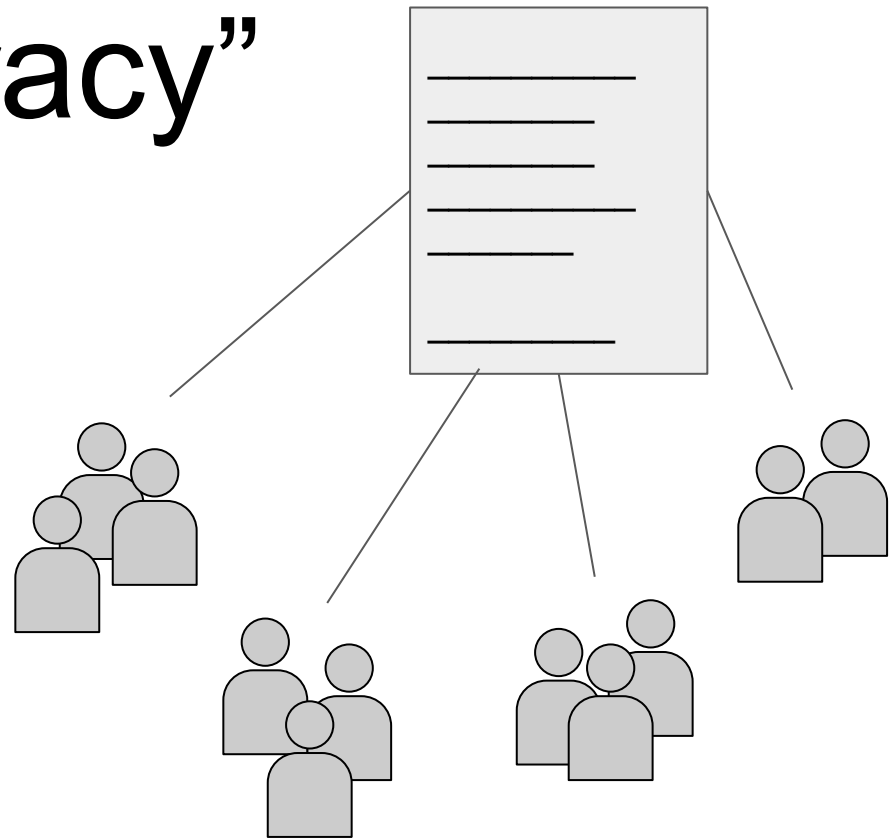
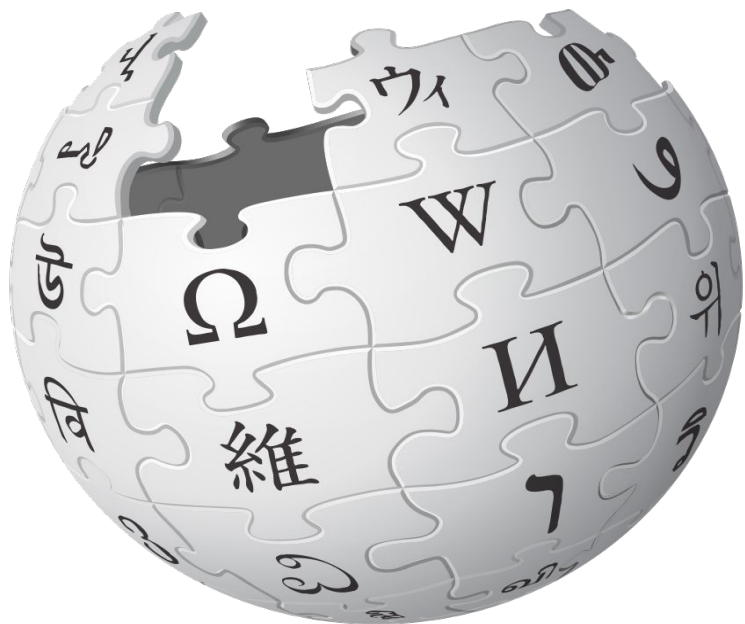
Let's teach them the basics of machine learning! Then, they'll understand what they are working with.



Precision-Recall
Cross-validation
Feature engineering
Hyperparameter optimization
Sample weighting

“Group Literacy”

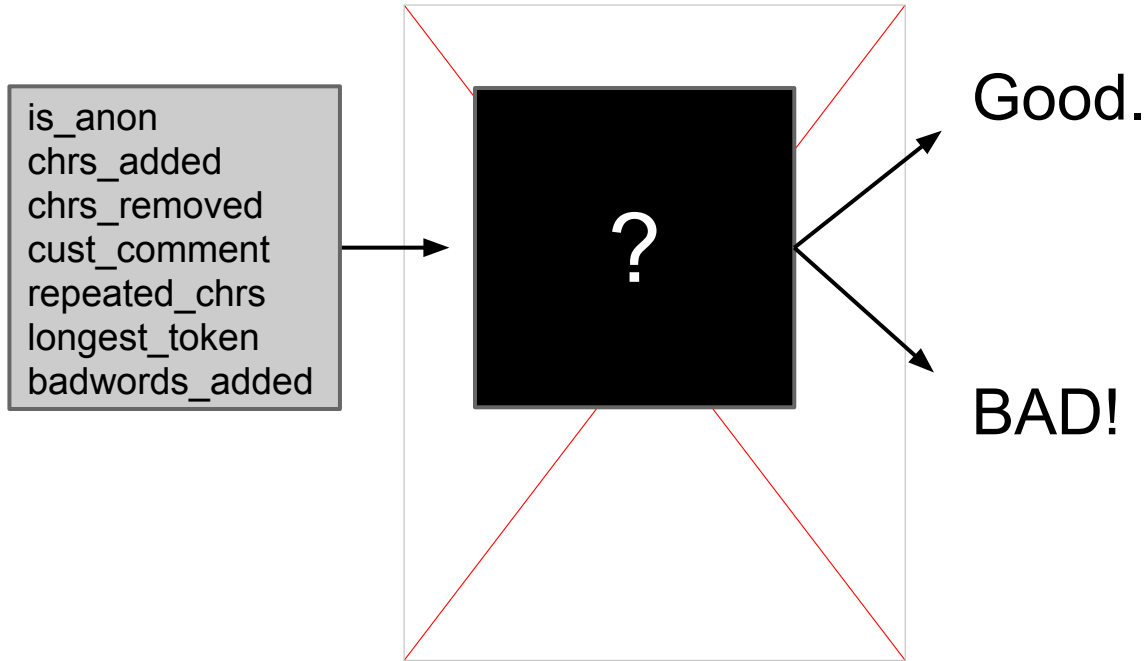
“Group Literacy”



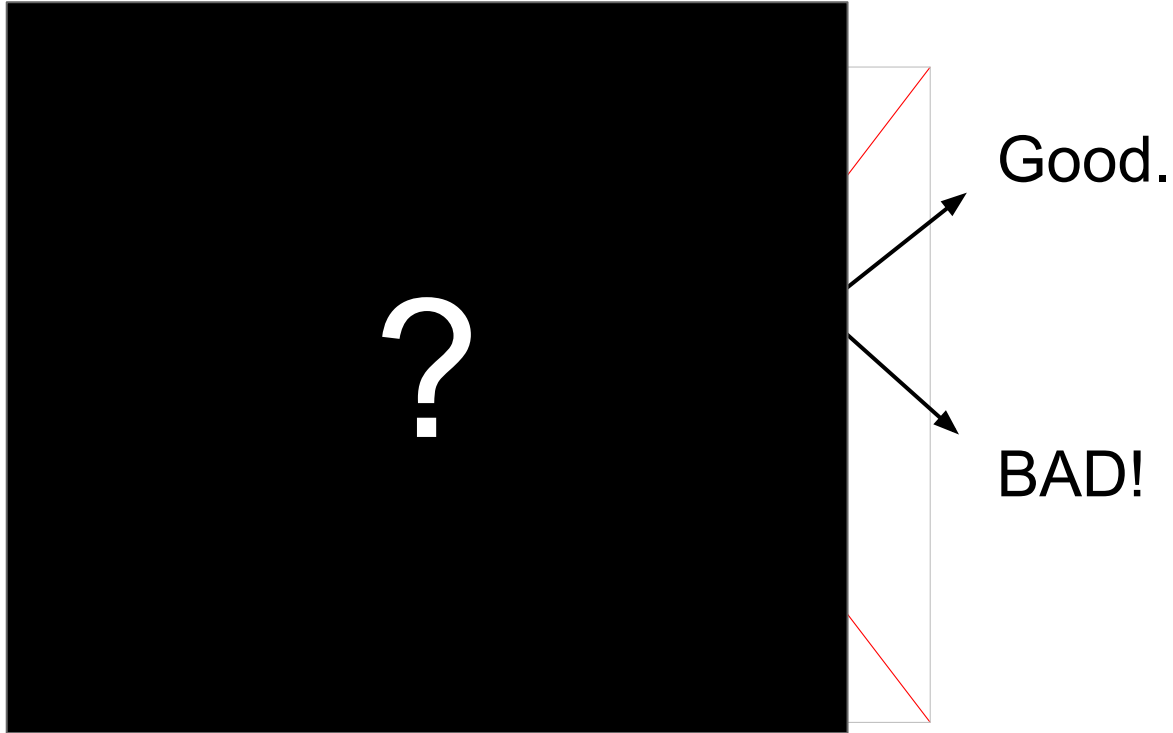
Part 3: The Pitch

Minimum Viable AI Governance

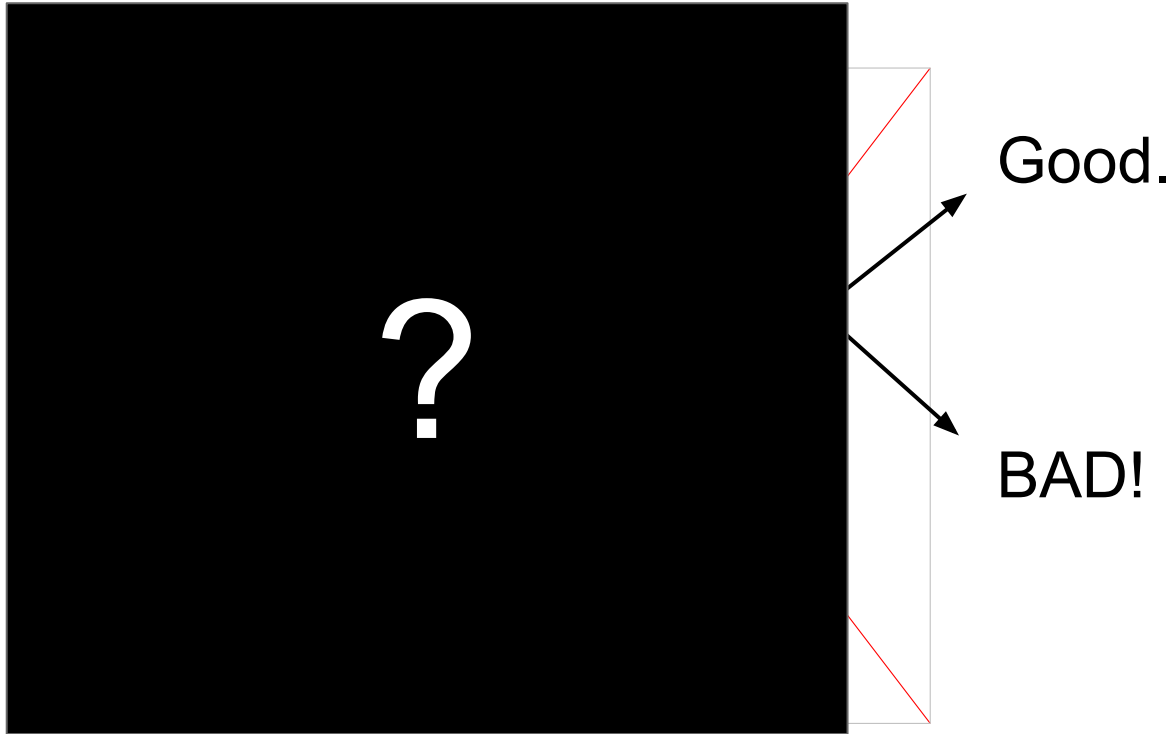
The machine classifier



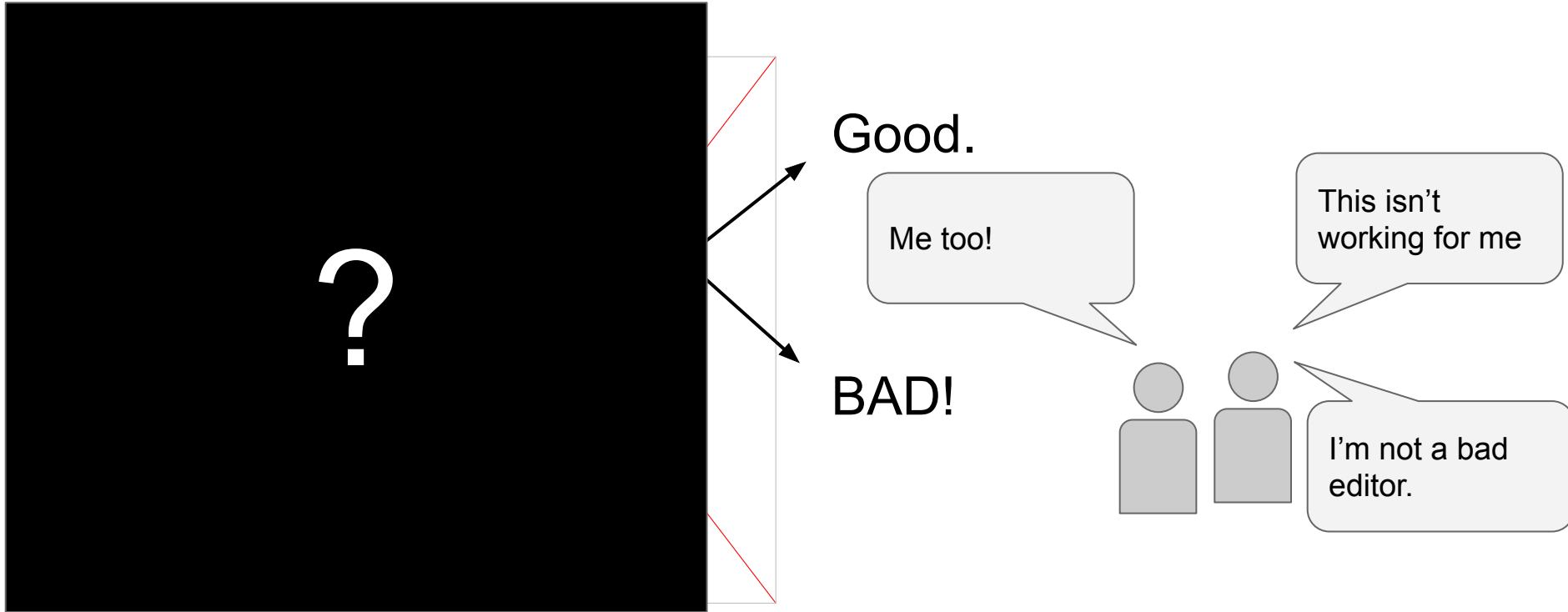
The machine classifier



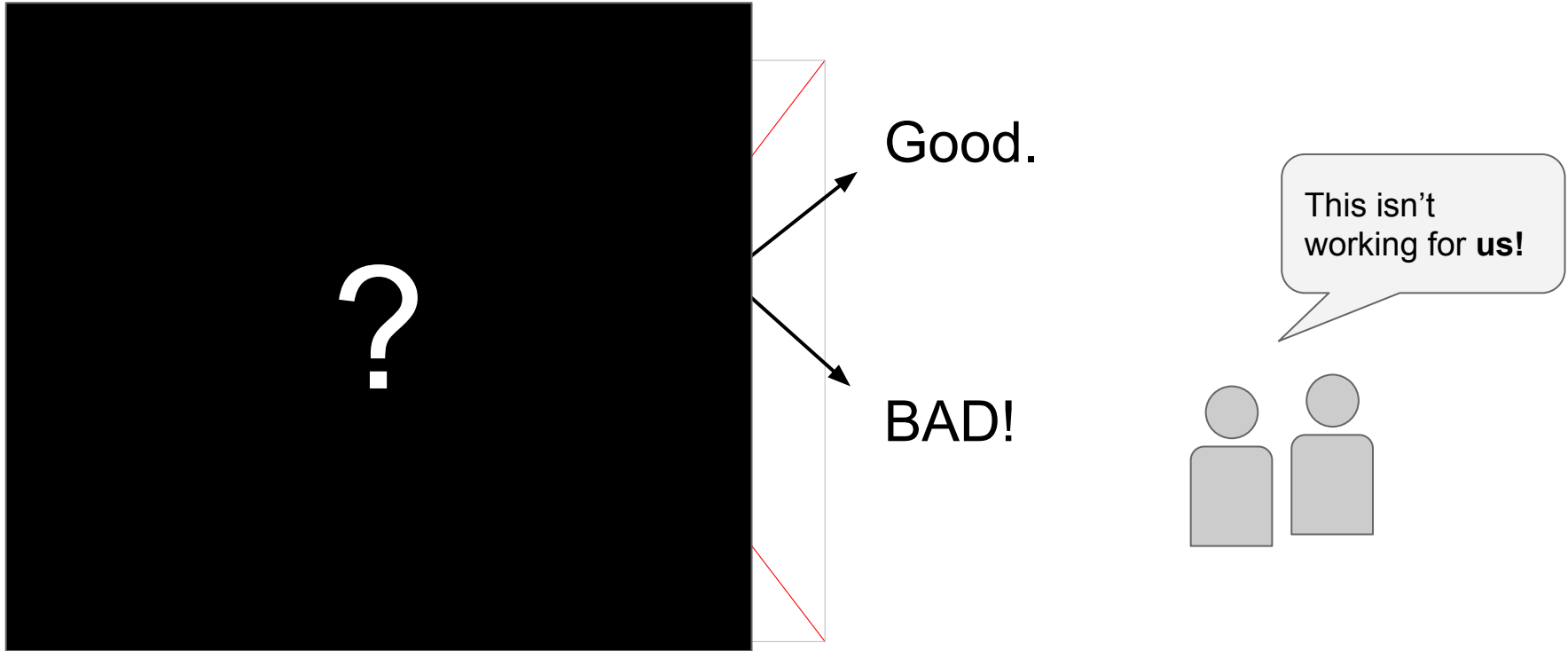
The machine classifier



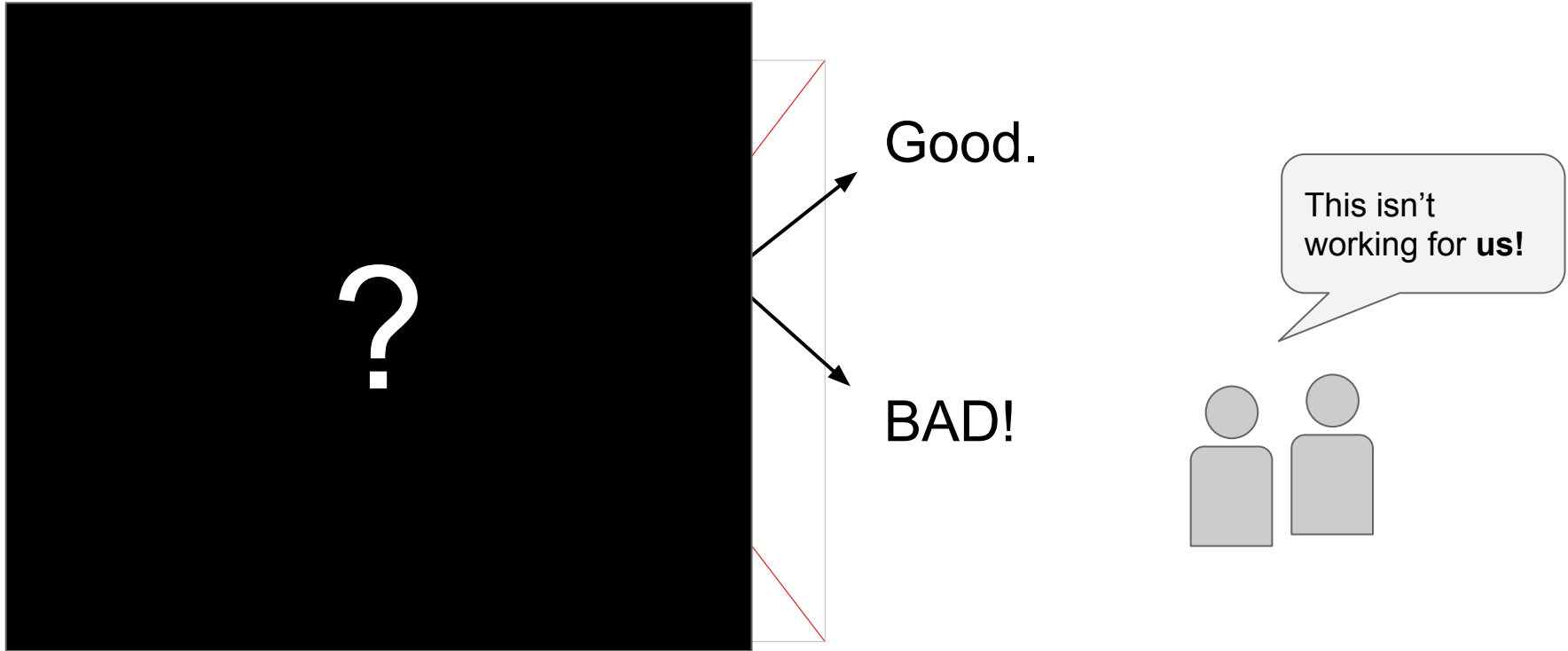
The machine classifier



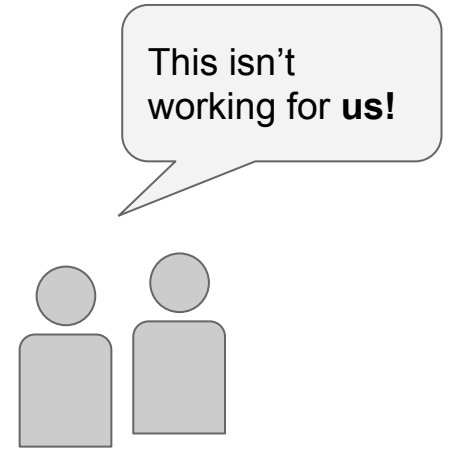
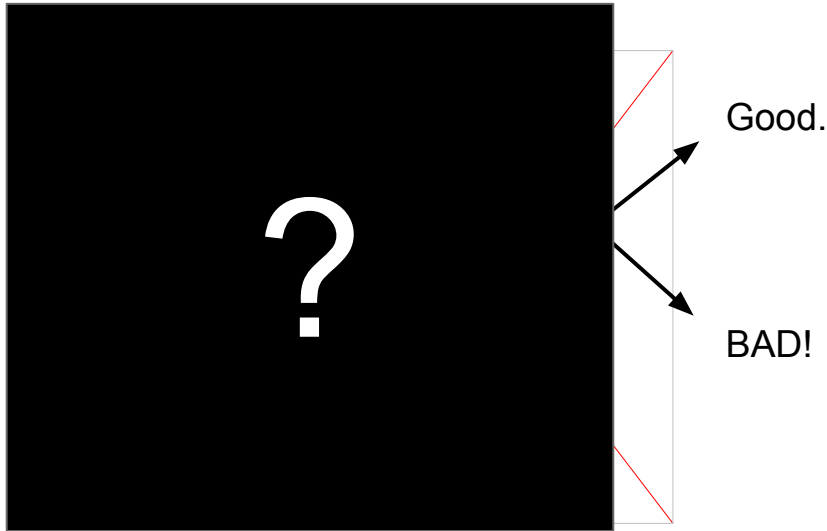
The machine classifier



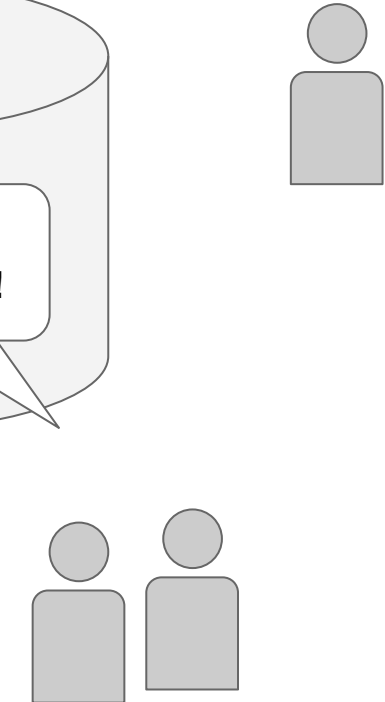
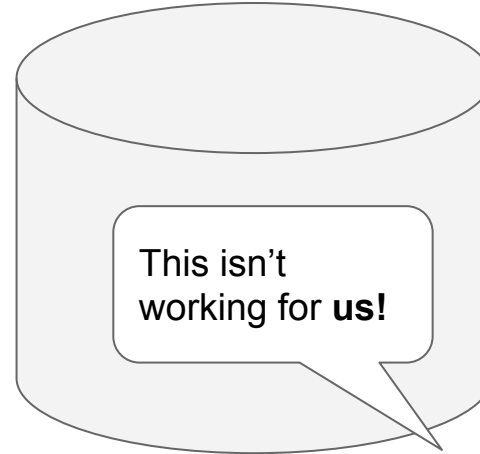
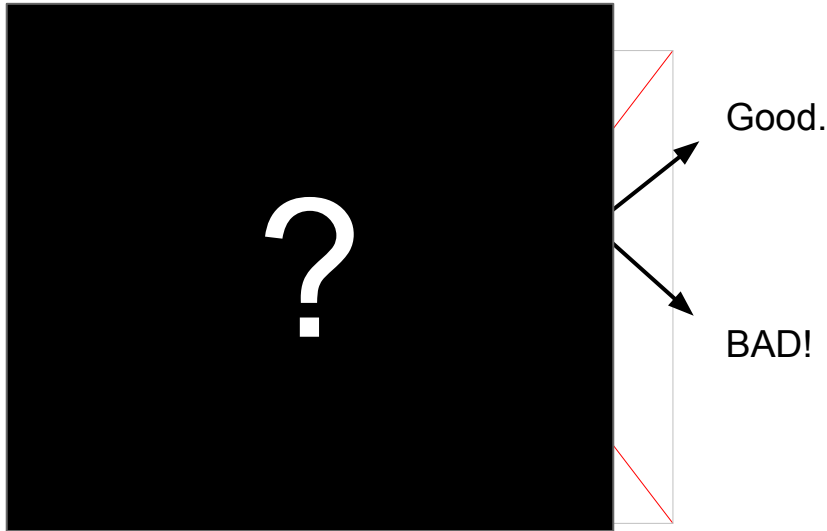
The machine classifier



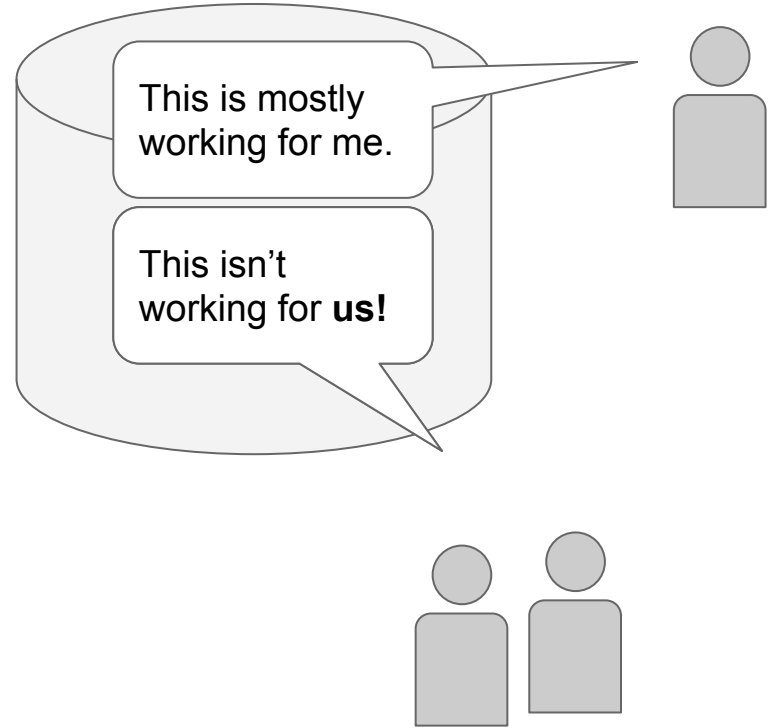
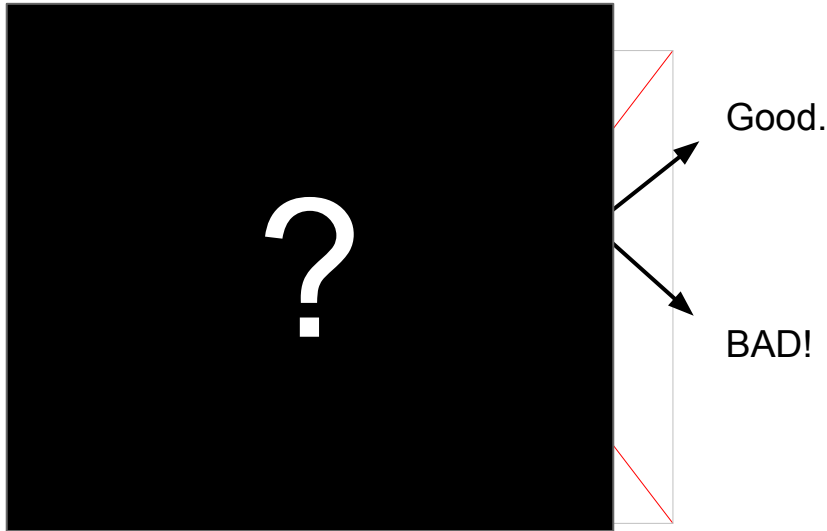
Minimum Viable Crowd-control



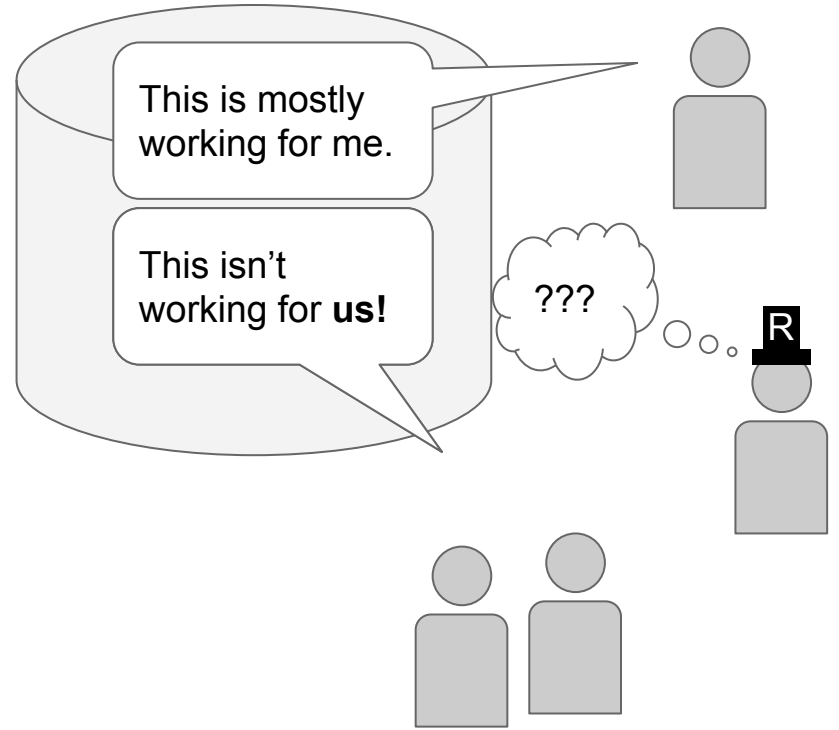
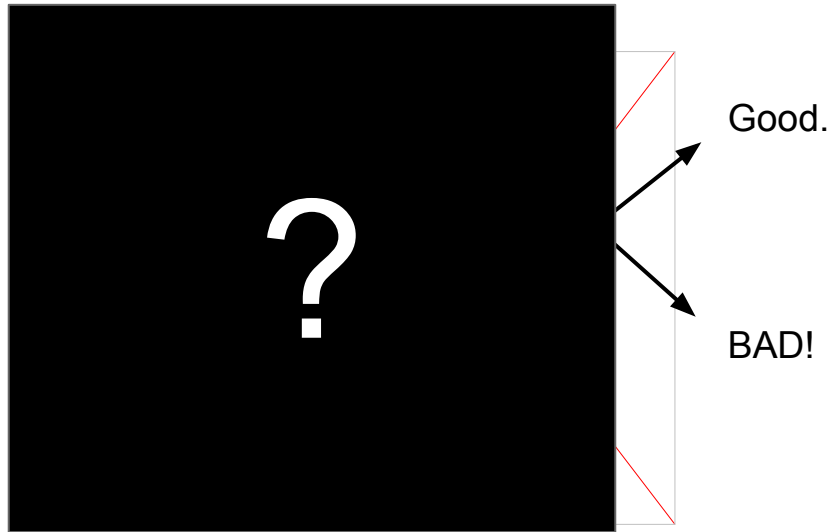
Minimum Viable Crowd-control



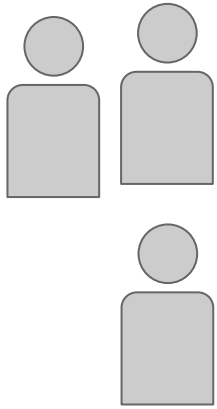
Minimum Viable Crowd-control



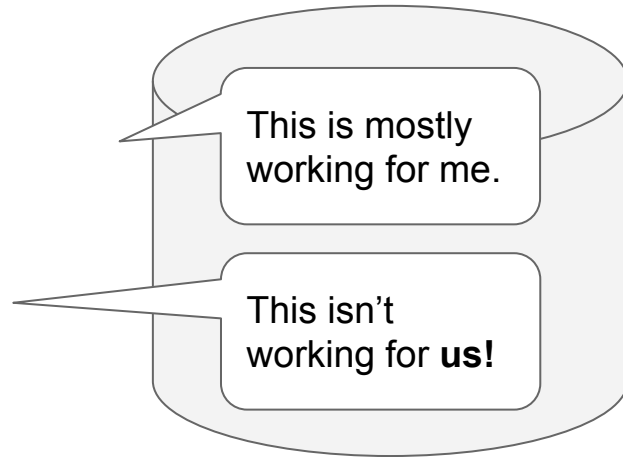
Minimum Viable Crowd-control



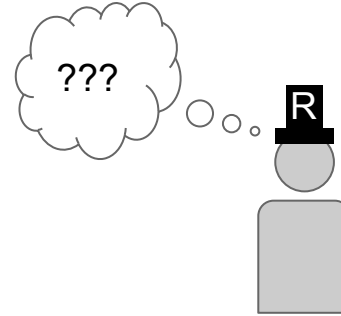
Minimum Viable Crowd-control



Users/stakeholders

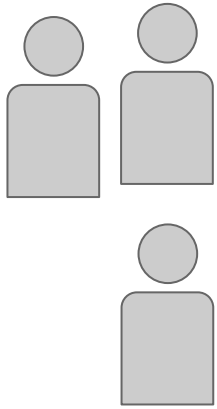


Repository of
successes/failures

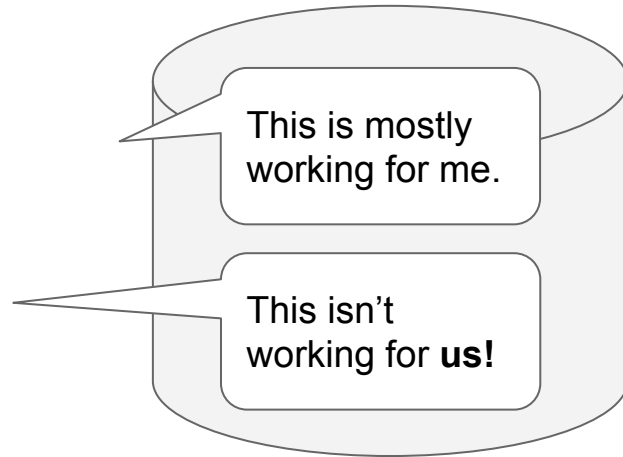


3rd party
researchers

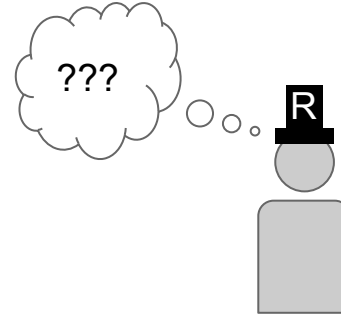
Minimum Viable Crowd-control



Users/stakeholders

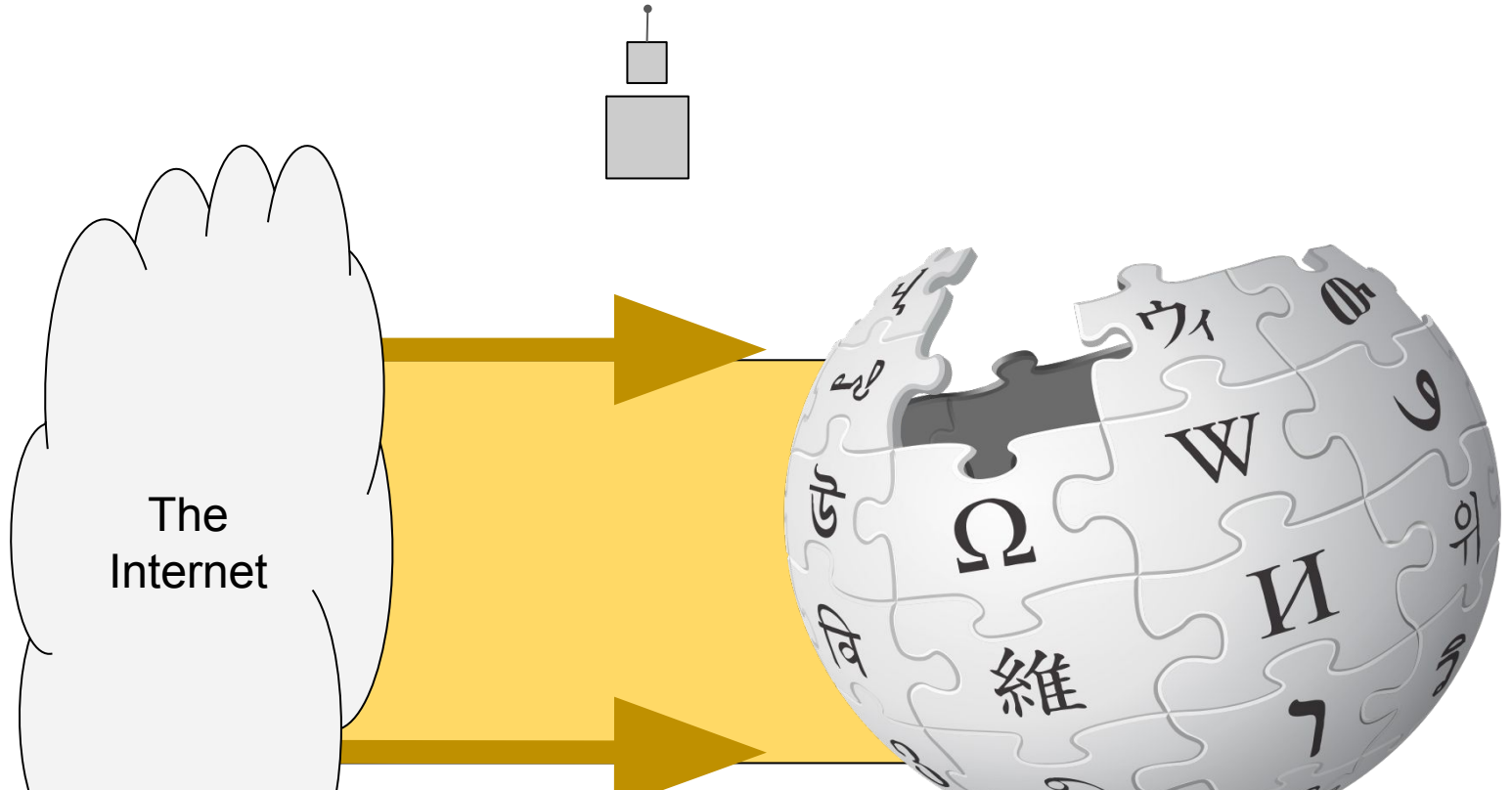


**Wiki Pages &
Templates**

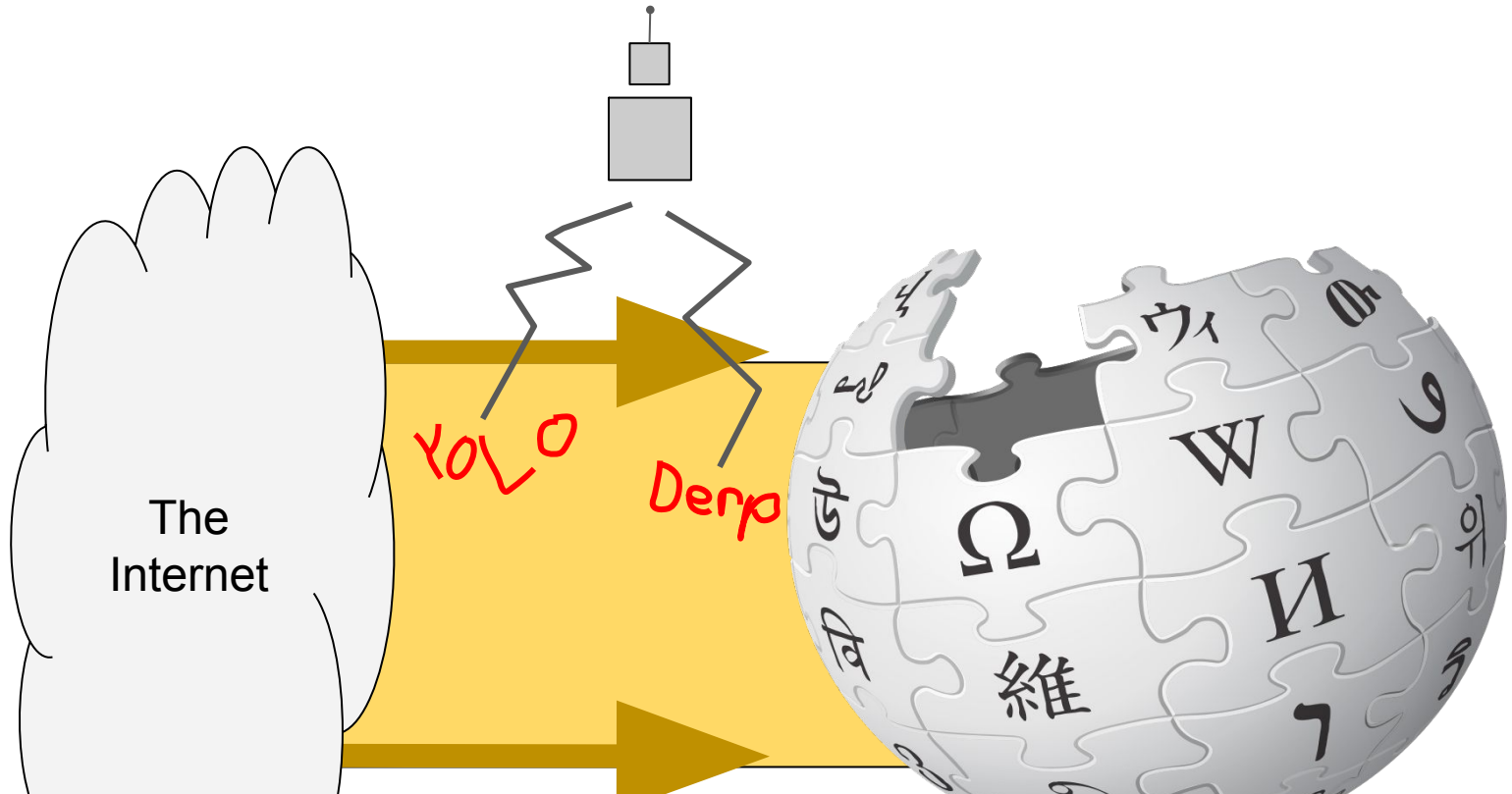


3rd party
researchers

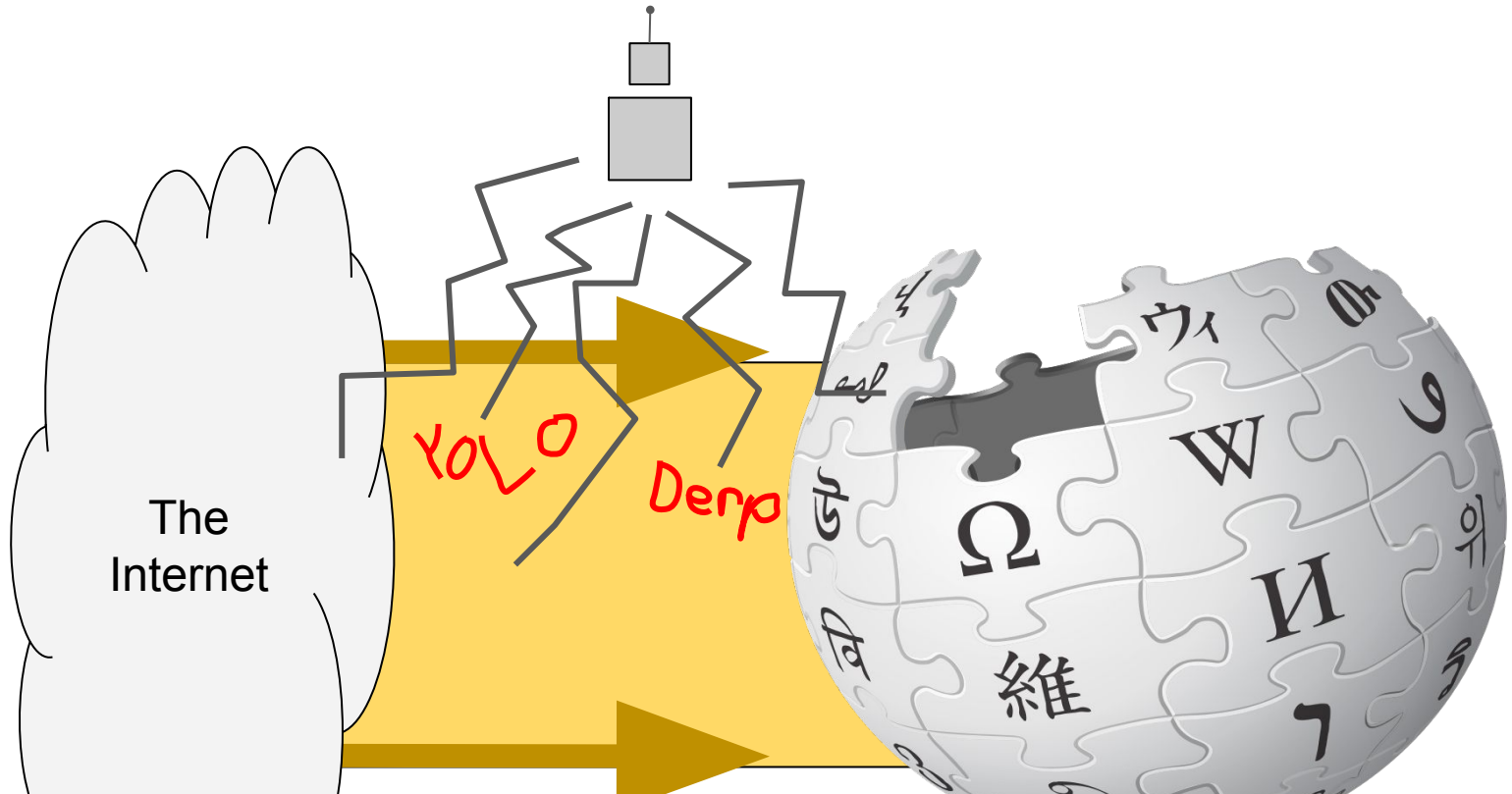
PatruBOT



PatruBOT



PatruBOT



Parada de PatruBOT [[editar código](#)]

Comunico que acabo de detener indefinidamente a PatruBOT. He visto que se han mantenido las quejas, y dado que anuncié esa posibilidad y que por ahora apenas si tengo tiempo de revisar (bastante por detrás) mi lista de seguimiento y de estar atento a IRC, creo que es lo más prudente. Notifico a [Mar del Sur](#) y [Ganimedes](#). Según mi tiempo disponible es posible que reconsidere en un futuro una reactivación con un nuevo algoritmo o si ORES facilita que sea más efectivo, en función también del interés que pueda tener la comunidad. En este sentido, leeré también aquello que se quiera comentar en este hilo. Saludos, - [José Emilio –jem– Tú dirás...](#)
09:34 27 mar 2018 (UTC)

Yo me desentendí hace bastante del asunto, pero gracias por el aviso. --Saludos. [Ganimedes](#) 12:14 27 mar 2018 (UTC)

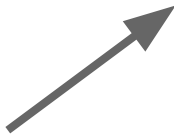
Una vez corregidos, al menos en parte, los problemas debería reactivarse, creo que es positivo. --[PePeEfe \(discusión\)](#) 13:11 27 mar 2018 (UTC)

Aunque cometiera (bastantes) errores, ayudaba muchísimo. De todos modos, si es la decisión de su operador, lo respetaré y aceptaré, por supuesto. Un bot antivandalismo me parece necesario en una Wikipedia tan visitada por vándalos que se empeñan en dejar su *huella* aquí. Espero que algo podamos hacer, porque con los reversiones o wikipedistas en general interesados en revisar los cambios recientes de los que disponemos actualmente no creo que tengamos suficiente. Un saludo cordial. --Fdo.: [Gonzalo P.M.G.](#) • 13:40 27 mar 2018 (UTC)

Considero que, a pesar de sus errores, el bot ayudaba mucho, ahora que ya no oesta activo los reversiones tendremos una tarea ardua. Espero que lo puedas activar de nuevo pronto. Saludos. --[Chico⁵¹²](#) 13:44 27 mar 2018 (UTC)

No me extraña que lo desactive por las constantes quejas que recibe. Algunas de estas quejas proceden de usuarios que llegan para editar un tiempo limitado, pero que se quejan de que se les han borrado datos, cuando en muchos casos lo que aportan son ediciones incorrectas. ¿Comete errores? Está claro, yo también, pero no me hago a la idea del trabajo que va a dejar de hacer. Los vándalos estarán encantados, y no me extrañaría que alguno fuese precisamente el que genera quejas para conseguir esto. Un saludo. --[vanbasten_23 \(discusión\)](#) 13:55 27 mar 2018 (UTC)

No extrañándome la decisión, que comprendo perfectamente, solo nos queda distribuirnos el trabajo lo mejor que podamos. Yo tengo marcas de seguimiento en todos los días y meses del año (efemérides), que son las páginas que patrullo continuamente. Supongo que hay otras formas de patrullar por intereses (no me veo patrullando deportes, por ejemplo). ¿Cómo se pueden implementar listas de patrullaje por categorías?. Saludos,--[Jmrebes \(déjame un mensaje aquí\)](#) 14:02 27 mar 2018 (UTC)





Wikipedia:Mantenimiento/Revisión de errores de PatruBOT/Análisis

< Wikipedia:Mantenimiento · Revisión de errores de PatruBOT



Esta página está destinada a analizar una muestra aleatoria de las contribuciones de PatruBOT.

Muestra aleatoria [[editar código](#)]

100 [[editar código](#)]

- 00:39 10 feb 2018 (+283)** [Berenguer de Cruïlles](#) (Revertidos los cambios de [95.16.136.45](#) a la última edición de TheRichic)-Revirtió borrado de texto y una parte de una plantilla - Gani
- 02:50 10 feb 2018 (-142)** [Unicanal \(Paraguay\)](#) (Revertidos los cambios de [181.120.150.204](#))- revirtió agregado de texto sin fuentes - Gani
- 02:59 10 feb 2018 (-40)** [Televisión digital terrestre en Paraguay](#) (Revertidos los cambios de [181.120.150.204](#) a la última edición de 2800:810:55D:17C9:957F:7965:2D6C:17B4)- revierte vandalismo - Gani
- 06:01 10 feb 2018 (+437)** [Sharknado 2: The Second One](#) (Revertidos los cambios de [179.41.146.136](#) a la última edición de FrescoBot)- revierte vandalismo - Gani
- 07:05 10 feb 2018 (-407)** [Colorina \(telenovela peruana\)](#) (Revertidos los cambios de [181.66.165.246](#) a la última edición de 179.7.208.250)- revierte edición que cambia apellidos, sin fuentes (posible vandalismo) - Gani
- 10:53 10 feb 2018 (-26)** [Configuración electrónica](#) (Revertidos los cambios de [83.56.97.82](#))- retira adición de cuenta de Twitter (posible spam) - Gani
- 11:08 10 feb 2018 (-613)** [PAW Patrol](#) (Revertidos los cambios de [190.46.1.9](#))- revierte correcciones de ortografía y adición de texto sin fuentes - Gani
- 13:36 10 feb 2018 (-156)** [Jueves Lardero](#) (Revertidos los cambios de [79.159.47.242](#))- revierte corrección menor y elimina adición de texto sin fuentes - Gani

Índice [[ocultar](#)]

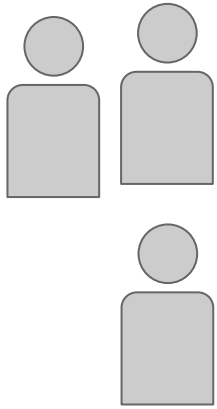
- Muestra aleatoria
 - 100
 - 200
 - 300
 - 400
 - 500
 - 600
 - 700
 - 800
 - 900
 - 1000
- Reservas
- Origen de los datos
 - Estadísticas globales
- Análisis de los datos
- Conclusiones

How often does PatruBot make a mistake?

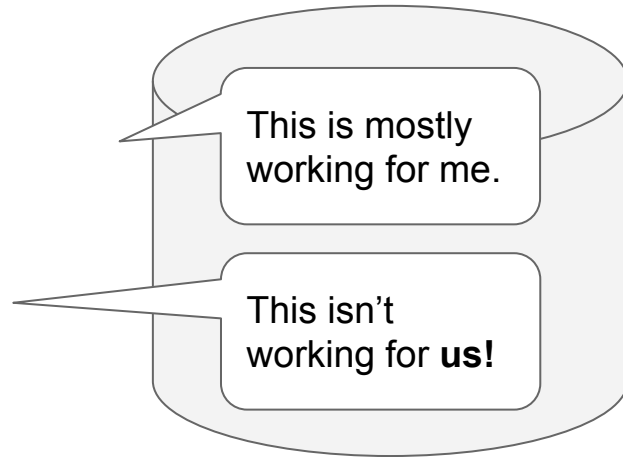
What kind of mistakes is it making?

Are these mistakes OK with us?

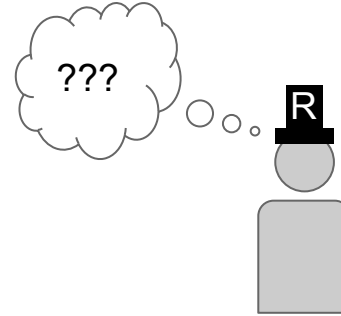
Minimum Viable Crowd-control



Users/stakeholders



“Jade”



3rd party
researchers

If it works for Wikipedia...

If it works for Wikipedia...

... maybe it'll work for Twitter, Google,
Facebook, Snapchat, etc.

If it works for Wikipedia...

... maybe it'll work for Twitter, Google, Facebook, Snapchat, etc.

And maybe it's a good candidate for policy.

Thank you!

Aaron Halfaker

Principal Research Scientist, Wikimedia Foundation

Think big. Measure what you can. Build better technologies.



About me

Hi. I'm Aaron Halfaker. I'm a scientist. See [projects](#) and [publications](#) below. I've been a Wikipedian since 2008. I mostly build tools

My work

My job is to build understanding about and support for the socio-technical fabric of the Wikimedia movement. I tend to focus on

Contact me

- E-mail: ahalfaker@wikimedia.org
- Website: <http://halfaker.info>
- Twitter: <http://twitter.com/halfak>